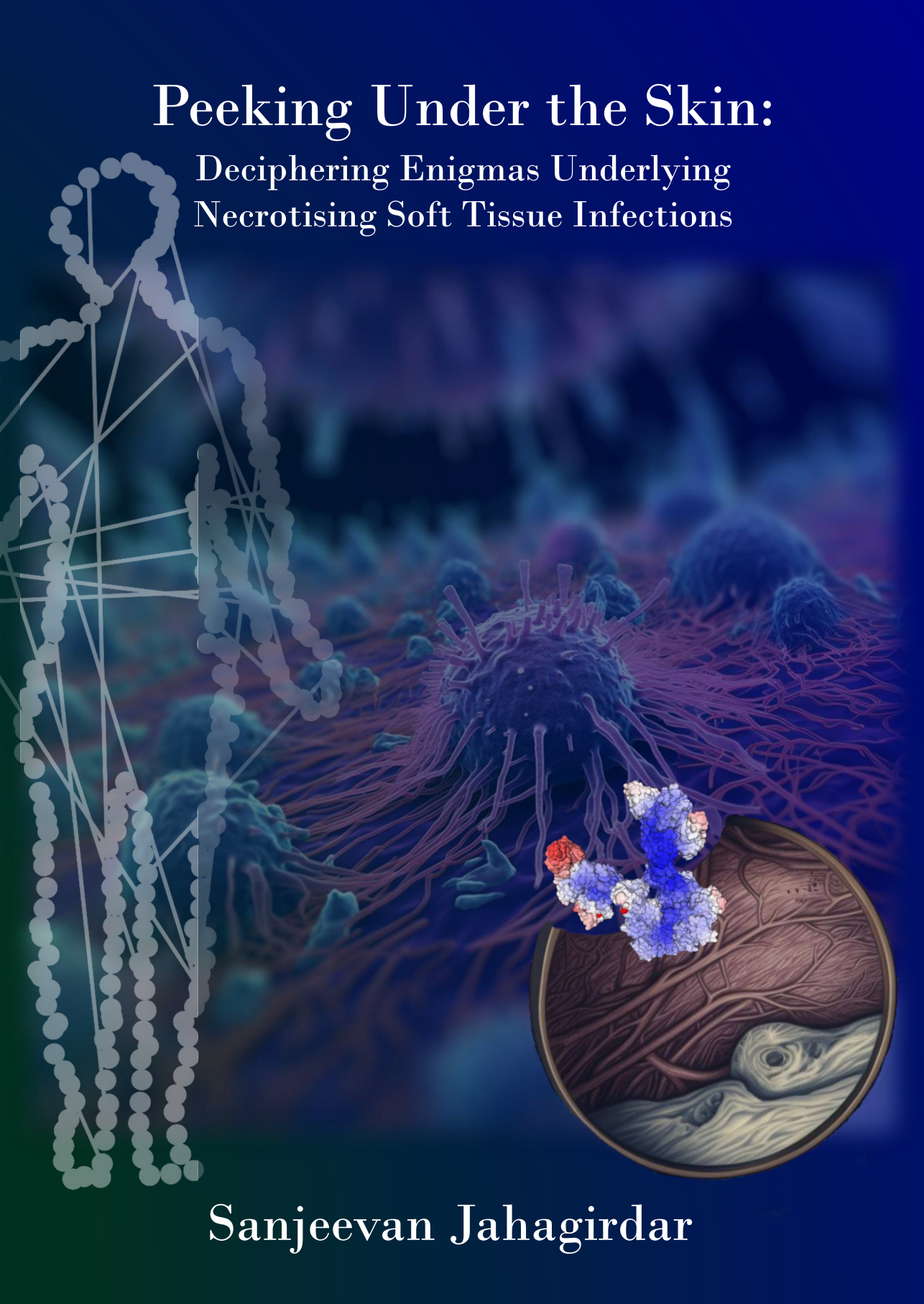
Peeking Under the Skin:

Deciphering Enigmas Underlying Necrotising Soft Tissue Infections



Sanjeevan Jahagirdar

Propositions

1. All models are right; however, subjectively defined features, information, and interpretations make them useful.
   (this thesis)

2. Replicating exploratory models does not turn them into predictive models.
   (this thesis)

3. The priority of scientific institutions needs to reside with illuminating stakeholders on the comprehension of scientific uncertainty.

4. Benevolent plausible deniability is dogmatically utilised in scientific discourse.

5. Large-scale scientific change is aloof as long as we measure scientific results instead of the scientific process.

6. The barrier to entry for open and reproducible science needs to decrease significantly as doctorates hold a monopoly neither on critical nor on divergent thought processes.

7. Science communication in social discourse is conflated with the diluting of complexities, nuances, and uncertainties of the acquired knowledge.

8. Divided we stand, united we fall.

Propositions belonging to the thesis, entitled

Peeking Under the Skin :
Deciphering enigmas underlying necrotising soft tissue infections

Sanjevan Jahagirdar
Wageningen, 16 October 2023

Peeking Under the Skin: Deciphering Enigmas Underlying Necrotising Soft Tissue Infections

Sanjeevan Jahagirdar

Thesis committee

Promotor
Prof. Dr Vitor A.P. Martins dos Santos
Personal chair, Bioprocess Engineering Group
Wageningen University & Research

Co-promotor
Dr Edoardo Saccenti
Assistant professor, Laboratory of Systems and Synthetic Biology
Wageningen University & Research

Other members
Prof. Dr Ellen Kampman, Wageningen University & Research
Prof. Dr Age K. Smilde, University of Amsterdam
Prof. Dr Marieke E. Timmerman, Rijkuniversiteit Groningen
Dr Lionel Blanchet, Genmab, Utrecht

This research was conducted under the auspices of VLAG Graduate School (Biobased, Biomolecular, Chemical, Food, and Nutrition sciences).

Peeking Under the Skin: Deciphering Enigmas Underlying Necrotising Soft Tissue Infections

Sanjevan Jahagirdar

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 16 October 2023
at 4 p.m. in the Omnia Auditorium.

Sanjevan Jahagirdar
Peeking Under the Skin:
Deciphering Enigmas Underlying Necrotising Soft Tissue Infections,
390 pages.

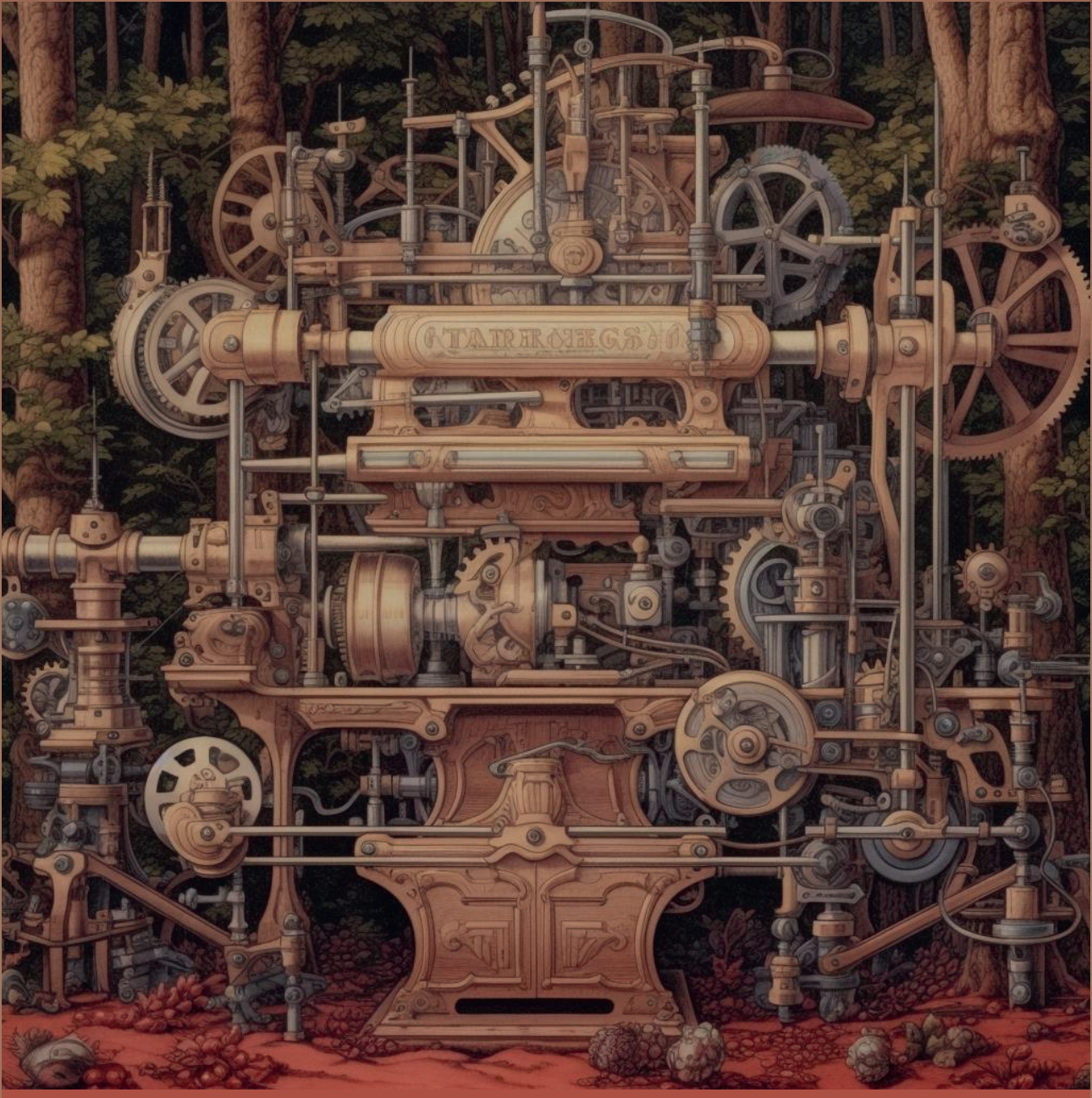
PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN: 978-94-6447-818-1
DOI: 10.18174/635496

To those who consciously choose to fight
for their personal rights and freedoms
that many others take for granted.

Contents

Chapter 1



Sanjeevan Jahagirdar[1]

Turn to page 377 for author affiliations

Introduction: Why this matters?

early more are bred than can survive, stated Charles Darwin in his joint essay with Alfred Russel Wallace while probing the fundamental question of the authority determining the selective survival of organisms in the long run **(Darwin and Wallace 1858)**. The essay demonstrated the evidence for evolution by natural selection and elaborated on the fundamental role death of an organism played in this process. He additionally expressed his perspective that nature was in a constant state of war, every individual/organism with another, or with external nature **(Darwin 2017)**. Today evolution by natural selection is a metaphorical spine with which all hypotheses in biology are connected. However, it is self-evident that the only reason an organism's demise plays such a crucial role in this process is that throughout the Earth's ∼3.9 billion year organic history **(Betts et al. 2018)**, every cell, organism, and species has spent their energy transpiring a state of biological animation. Ironically, apart from instances of inter- and intra-species violence, the relentless struggle for survival among viruses, bacteria, fungi, plants, and animals constitutes a significant contributing factor to organismal mortality.

Indeed, parasitism is not exclusive to particular genera; instead, it functions as a survival strategy, representing yet another niche in the environment exploited by opportunistic species and evolutionary processes **(Poulin et al. 2000)**. The ability of these species to thrive and reproduce hinges on their capacity to usurp at least a portion of the resources required or produced by other species. These parasitic strategies, which can involve intricate inter-species interactions, are diverse and often highly complex **(Paul et al. 2003)**. Certain worms, for instance, can seize total control over an insect's body, thus forcing a behavioural change that suits the worm's life cycle **(Biron et al. 2005)**. Conversely, other parasites have the ability to manipulate rodents' perceptions, diminishing their fear of felines **(Tong et al. 2021)**.

This struggle for survival can either lead to a symbiotic mechanism or an evolutionary arms race. Viruses have been documented to infect larger viruses, which in turn infect more complex life forms **(La Scola et al. 2008)**. Bacteriophages (viruses that infect bacteria) have precipitated one of the most extensive and enduring biological arms races, with casualty figures often in the trillions, if not higher **(A. Stern et al. 2011)**. This relentless conflict has not only resulted in intricate offensive and defensive systems (which we exploit in synthetic biology) **(Makarova et al. 2011)**, but has also impacted the global nutrient cycles **(Suttle 2007)**, climate **(Fuhrman 1999; Suttle 2007)**, the evolution of the biosphere **(Comeau et al. 2005)**, and the evolution of virulence in human pathogens **(Brussow et al. 2004)**.

From a human standpoint, certain manifestations of these factors are recognised as diseases. The concept of "disease" has continuously been a subject of intense debate,

entailing scientific, philosophical, and societal dimensions. Diseases, being physical or psychological abnormalities, are diverse in their sources and impacts, thereby challenging a universal definition. Attempts from defining diseases have ranged from viewing the concept of disease as unnecessary, focusing instead on desired medical interventions **(Hesslow 1993)**, to establishing definitions of disease based on naturalist or normativist theories **(Powell et al. 2019)**. Naturalist theories have been scrutinised for potentially categorising human physiological diversity and neuro-diversity in the realm of disease **(Boorse 1977; Kingma 2010)**. Meanwhile, normativist characterisations, focusing on holistic and quality of life perspectives, grapple with the issue of pathologising behaviours that deviate from cultural norms **(Powell et al. 2019; Goosens 1980)**.

Narratives and stories also play a crucial role in the definition, perception and understanding of diseases. Research has shown that stories can strongly influence public perceptions, skew beliefs in line with the narrative, and even elicit strong emotional responses leading to cognitive alignment with the narrative's context **(Green et al. 2000)**. This effect of narratives has been leveraged in narrative medicine to provide a voice to patients' experiences and perspectives **(Charon 2008; Cartledge et al. 2020)**. I would like to guide the interested reader to the section *The disease dilemma* in **Chapter 11** for a further in-depth discussion on these concepts.

This thesis is a scientific inquiry using computational approaches to understand and explore the underlying mechanisms in necrotising soft tissue infections (NSTI). NSTIs are a predominantly bacterial disease associated with necrosis, sepsis **(Hua et al. 2022)**, a high mortality rate **(D. L. Stevens and Bryant 2017)**, and severe loss in the quality of life **(Suijker et al. 2020)**. This disease has often been colloquially termed as "flesh-eating disease" **(Davies 2001)**.

1.1 Necrotising Soft Tissue Infections

"The only way I could get him comfortable was to lay him on my chest facing me." wrote Lucy Dove recounting the story of her 1-year-old son's necrotising soft tissue infection **(Cartledge et al. 2020)**. She further added, "On his left side approximately halfway up, I pressed Frankie cried out in pain. I worked out there was an area of approximately 2 × 2 inches that was very painful for Frankie when I touched it." Lucy's story describes the harrowing experience of a 1-year-old boy, Frankie, who exhibited flu-like symptoms and developed severe pain and swelling in various parts of his body. Despite initial misdiagnosis and delayed treatment, Frankie was eventually diagnosed with NSTI. Extensive surgical intervention was required to remove infected tissues, and Frankie's condition remained critical for an extended period of time. Through multiple surgeries, skin grafts, and a long recovery process, Frankie survived but endured prolonged physical and emotional challenges, including recurrent infections, scarring, and the need for ongoing medical care **(Cartledge et al. 2020)**. Photos of Frankie showing his harrowing experience are shown in Figure 1.1.

1.1.1 Why this matters?

Necrotising soft tissue infections (NSTIs) are a group of infections that are characterised by widespread tissue destruction in any layer of the soft-tissue compartment

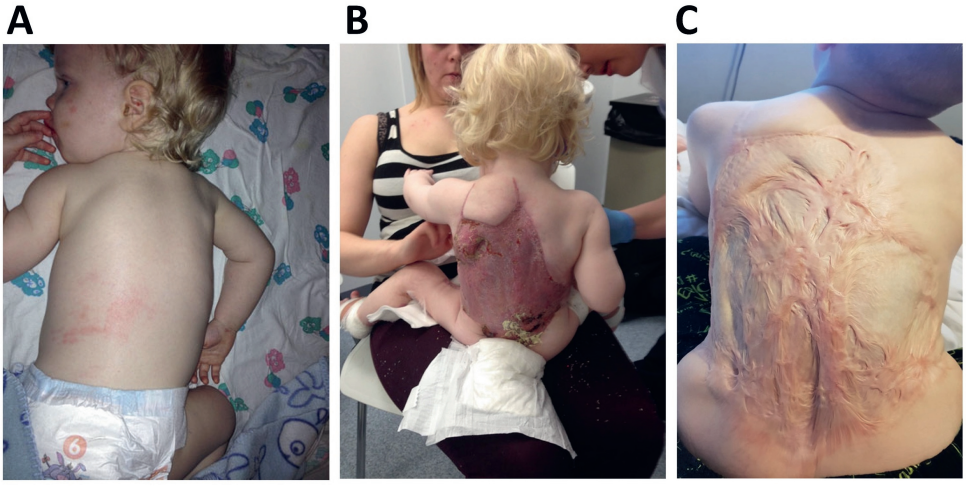


Figure 1.1: A series of images documenting Frankie's harrowing experience **(Cartledge et al. 2020)**. The images are used unmodified from **(Cartledge et al. 2020)** with permission from Springer Nature under the license number 5583650779496. (**A**) Frankie's back the morning after being admitted. Clearly visible are swellings on his left side and changes to the skin colour of the surface. (**B**) Frankie's first dressing change on the ward. (**C**) Frankie in January 2020 **(Cartledge et al. 2020)**

ranging from the epidermis (top layer of skin) to deep musculature (deep back muscles) **(D. L. Stevens and Bryant 2017)**. These infections are often characterised by the friability of the superficial fascia (layer directly under the skin), dishwater-grey exudate (fluid inside tissue turns a grey, murky colour), and a notable absence of pus (a thick yellowish or greenish opaque liquid produced in infected tissue) **(D. L. Stevens and Bryant 2017)**. This subcutaneous tissue, fascia or muscle necrosis is associated with high morbidity and mortality **(Hua et al. 2022)**. NSTIs are relatively rare and the estimated incidences vary based on geographical areas. The estimated incidence rates of the condition range from 0.2 to 6.9 per 100,000 person-years **(Naseer et al. 2016; G. Glass et al. 2015; Bocking et al. 2017)**, with the highest reported rate of 15.5 per 100,000 person-years observed in Thailand **(Khamnuan et al. 2015)**. Nevertheless, it is important to note that rates in other regions may potentially be higher, and precise data may not be available at present **(Barupal et al. 2019)**. However, incidence rates estimates may not be very accurate due to the rarity of the disease, the absence of systematic reporting policies, and the use of multiple terms to describe the ailments such as necrotising fasciitis, hospital gangrene, Fournier's gangrene, Meleney's gangrene, synergestic gangrene, clostridial cellulitis, and necrotising cellulitis to name a few **(Hua et al. 2022)**. Despite the rarity of the disease, the occurrence of NSTI has been increasing over the last decades **(Oud et al. 2015; Das et al. 2011; Bodansky et al. 2020; Hedetoft, M. B. Madsen, et al. 2020; G. E. Nelson et al. 2016; Plainvert et al. 2012)** and Hua et al. claim that most physicians might see a minimum of one case of NSTI in their career **(Hua et al. 2022)**.

NSTIs possess a significant health burden both on the suffering individual and the

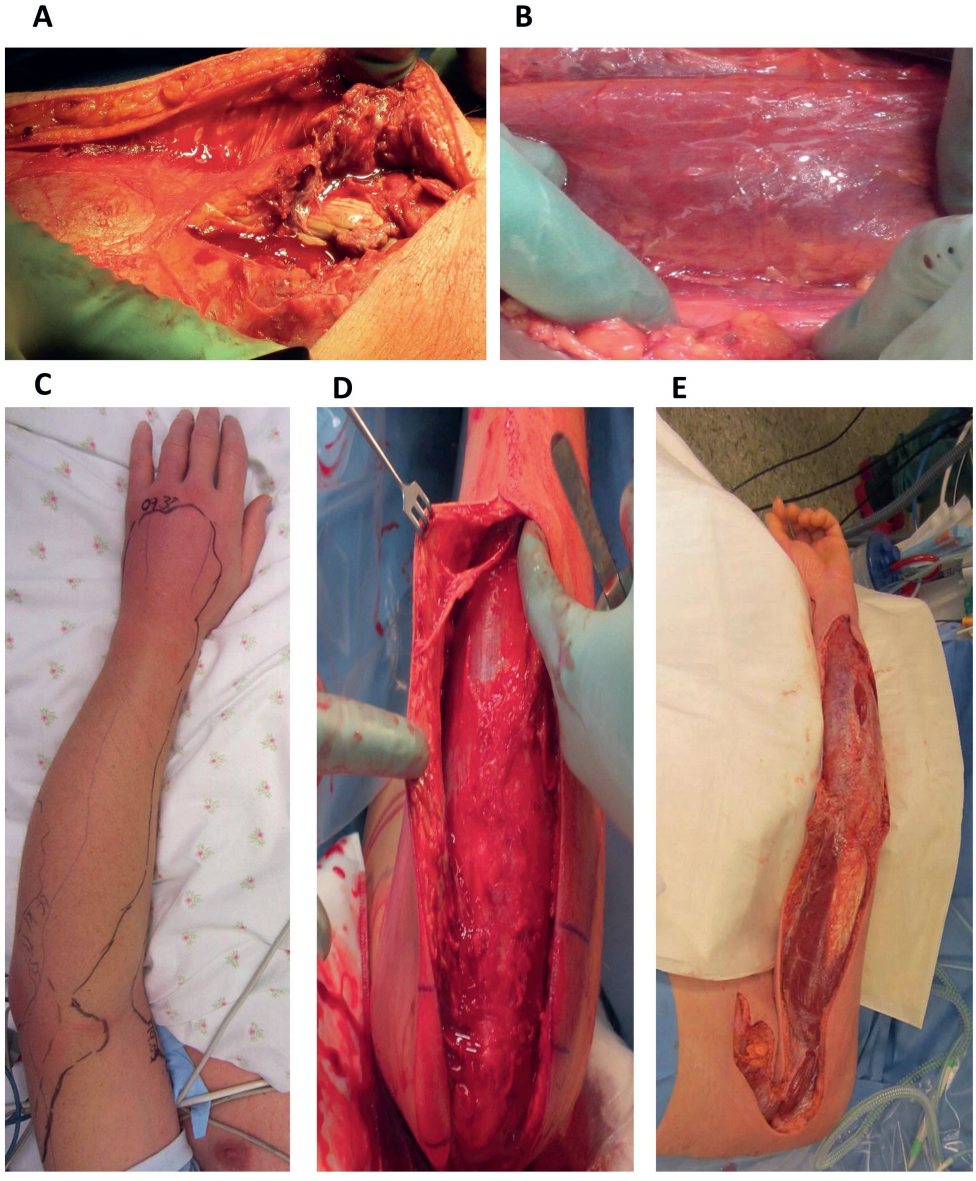


Figure 1.2: A series of images showing the tissue destruction in NSTI. The images are produced unmodified with permission from Springer Nature under the license number 5583660415667 **(Nedrebo et al. 2020)** (**A**) Necrotic muscle (parts of quadriceps) from a polymicrobial infection/anaerobic infection originating in the pelvic region. (**B**) Edema and dishwater-like fluid under the fascia after GAS infection in an immunocompromised patient. (**C**) GAS of the upper extremity: spread of erythema and swelling down to MCP joints. The picture was taken 12h after first admission. (**D**) GAS of the upper extremity: first surgery over the elbow and forearm. 7 × 20 cm of the skin removed due to necrosis. (**E**) GAS of the upper extremity: surgery after the second revision had to go below the shoulder **(Nedrebo et al. 2020)**.

society and its infrastructure **(A. K. May et al. 2009; Pham et al. 2009; M. B. Madsen, Skrede, et al. 2019)**. Mortality rates of NSTI patients have been reported to be around 23.5% **(A. K. May 2009)**, however, the mortality rate can range from 4% to 60% based on the time taken for diagnosis and surgical treatment **(Nawijn et al. 2020)**. The decrease in mortality rates over the decades also leads to an increase in the number of patients living with the consequences of NSTI **(Suijker et al. 2020)**. NSTI is also associated with severe long-term disability due to the treatment itself **(A. K. May et al. 2009)**. The treatment consists of surgical debridement of all necrotic tissue and administration of antibiotics which leaves the patients with amputations, extensive wounds, mutilating scars, and functional deficits **(Urbina et al. 2021; Suijker et al. 2020)**. The progression of the disease itself and the subsequent scars can both severely influence the patient's health-related quality of life **(Suijker et al. 2020)**. Figure 1.2 displays select photos of NSTI tissue destruction and surgical debridement.

1.1.2 Symptoms

NSTI presentation is greatly variable **(M. B. Madsen, Arnell, et al. 2020)**, often linked to breaches of the skin due to ulcers, trauma, or surgery, but can also occur without previous penetrating trauma **(McHenry et al. 1995; Wong, H.-C. Chang, et al. 2003; M. B. Madsen, Skrede, et al. 2019; B. Wilson 1952; Meleney 1924)**. Symptoms can range from pain and swelling at the site of infection to more systemic symptoms such as fever, tachycardia (heart rate over 100 bpm), hypotension (low blood pressure), and altered mental state **(Bisno et al. 2000)**. Having said that, Hua et al report that as many as 40% of NSTI cases may not have a fever **(Hua et al. 2022)**. NSTI patients present vague symptoms (nausea, vomiting, and diarrhea **(Bisno et al. 2000)**) and there is no particular symptom that is definitive of NSTI, making the diagnosis very challenging **(M. B. Madsen, Arnell, et al. 2020)**. This is further compounded by the fact that a quick diagnosis followed by surgery has been linked to a reduced mortality rate **(Nawijn et al. 2020)**. Moreover, suspicion of NSTI is often developed with measurements taken at a time when patients are already in the hospital. Be that as it may the case, individuals may experience various symptoms before admission to the hospital that may play a critical role in the timely diagnosis **(Erichsen Andersson et al. 2018)**. Comorbidities are common among NSTI patients and they may include diabetes, cardiovascular disease, chronic kidney and liver disease, and other immunocompromising diseases **(Boyer et al. 2009; Sudarsky et al. 1987)**. In the INFECT cohort, comorbidities were more common among non-streptococcal than streptococcal cases **(Bruun, Rath, et al. 2021)**. Some of these conditions may increase the risk of NSTI, but their definitive association with NSTI occurrence remains unclear **(Goh et al. 2014)**.

1.1.3 Diagnosis

The early diagnosis of NSTI has proven to be challenging due to several factors including vague symptoms, heterogeneous patient groups, lack of specific diagnostic tools, and other factors detailed above **(T. Chan et al. 2008)**. It has been reported that as many as 50% of the patients are initially misdiagnosed **(A. K. May 2009; Goh et al. 2014)**. The gold standard for confirming a NSTI diagnosis is a surgi-

cal exploration that reveals specific inoperative findings **(Hua et al. 2022)**. Other methods used for diagnosing NSTIs encompass imaging techniques, laboratory measures and investigations, and microbiological diagnoses **(D. L. Stevens and Bryant 2017; Hua et al. 2022)**. However, Hua et al. warn that urgent surgical exploration should always supersede other methods including imaging techniques, given the latter's limited sensitivity in differentiating NSTI from non-necrotising soft tissue infections (Non-NSTI) **(Hua et al. 2022)**. Potentially several imaging options could be utilised in the diagnostic process **(M. B. Madsen, Arnell, et al. 2020)** including plain radiographs **(Tso et al. 2018)**, ultrasonography **(Yen et al. 2002)**, computerised tomography **(Leichtle et al. 2016)**, and magnetic resonance imaging (MRI) **(Kim et al. 2011; Rahmouni et al. 1994)**. MRI has proven to be useful, particularly in limb infections **(Kim et al. 2011; Rahmouni et al. 1994)**. In the clinic, the classical manifestations of NSTI include soft-tissue edema (75% of the cases), erythema (72%), severe pain (72%), tenderness (68%), fever (60%), and skin bullae (38%) **(McHenry et al. 1995; D. L. Stevens and Bryant 2017)**. Laboratory Risk Indicator for Necrotising Fasciitis (LRINEC) is a score developed retrospectively by Wong et al. based on the serum C-reactive protein, leukocyte count, haemoglobin, sodium, creatinine, and glucose values **(Wong, Khin, et al. 2004)**. However, the utility of LRINEC has been called into question due to studies reporting low sensitivity and specificity **(M. B. Madsen, Skrede, et al. 2019; Hsiao et al. 2020; Fernando et al. 2019; Bechar et al. 2017; Neeki et al. 2017; Putnam et al. 2016)**. Katz et al. showed high accuracy in predicting 30-day mortality in NSTI patients using 16 clinical parameters in their decision support system, however, this is yet to be externally validated **(S. Katz et al. 2022)**. Although the initial treatment is not influenced by specific microorganisms present at the time of admission, microbiological diagnosis can be used to primarily identify the type and nature of infection-causing microorganism(s) **(D. L. Stevens and Bryant 2017; Hua et al. 2022)**.

1.1.4 Classifications

Accurate identification of the microbial aetiology is crucial for appropriate prognostic information and optimal antibiotic therapy **(D. L. Stevens and Bryant 2017; Anaya et al. 2005; K.-F. Huang et al. 2011; M. B. Madsen, Skrede, et al. 2019)**. However, empirical antibiotic treatment is often inadequate in 90% of the patients due to the substantial knowledge gap in microbial aetiology of NSTIs, caused by the rarity of the condition, frequent involvement of anaerobic microbes, and diagnostic challenges **(Marwick et al. 2011; Khamnuan et al. 2015; Childers et al. 2002)**. Historically NSTIs have been classified based on their anatomical locations resulting in numerous clinical definitions **(Skrede et al. 2020)**. The utility of microbiology in this context has been increasingly recognised, with new methods of pathogen identification and resistance profiling emerging to support tailored therapeutic measures **(Skrede et al. 2020)**. Modern approaches have tended to classify NSTIs based on their microbial aetiologies. A pragmatic classification approach is presented in Figure 1.3. In general, NSTIs have been classified into the following sub-groups based on microbial aetiologies **(Skrede et al. 2020; D. L. Stevens and Bryant 2017; Hua et al. 2022)**:

- **Type I**: Polymicrobial infections (mixed aerobic and anaerobic organisms)

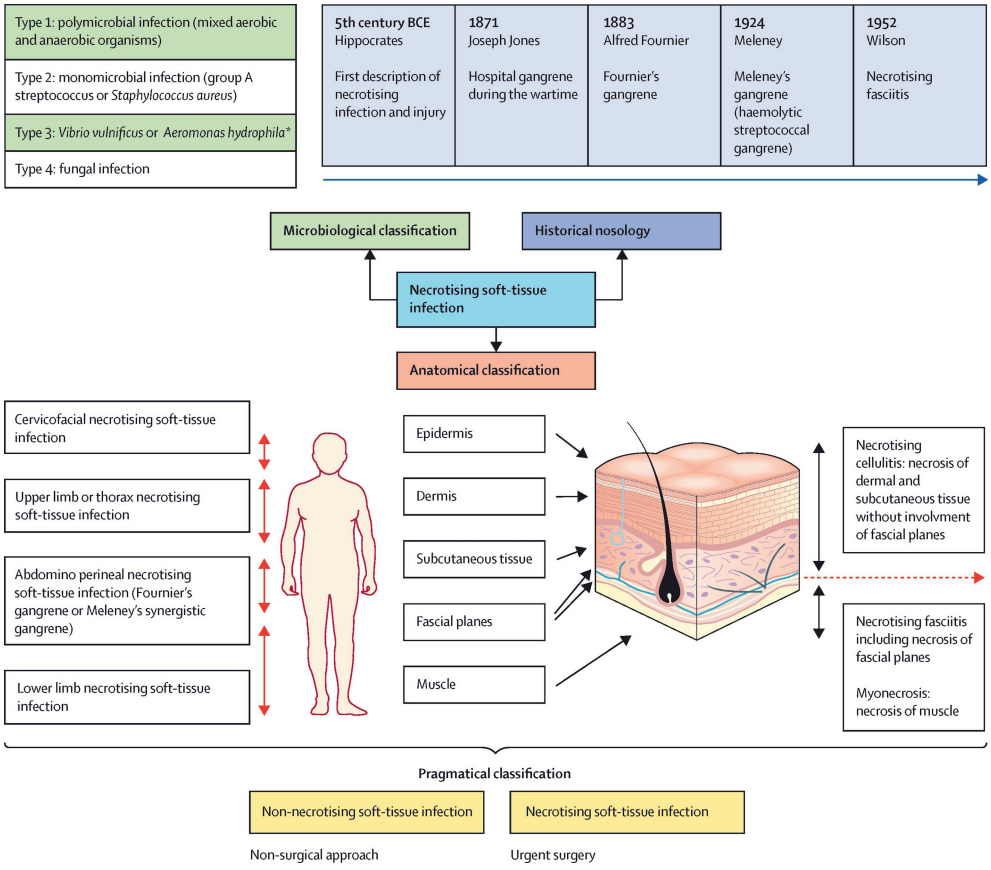


Figure 1.3: Pragmatic classifications of NSTI from **(Hua et al. 2022)** published here unmodified with permission from Elsevier under the license number 5583651359700. Specifically highlighted is the pragmatic classification of non-NSTI. Non-NSTI are the cases where NSTI is suspected, however, no necrosis is found upon surgical evaluation.

- **Type II**: Monomicrobial infections (*Streptococcus pyogenes*, *Streptococcus dysgalactiae*, & *Staphylococcus aureus*)
- **Type III**: *Vibrio vulnificus* or *Aeromonas hydrophila* or clostridial or monomicrobial Gram-negative infections
- **Type IV**: Fungal infections.

1.1.5 Microbial aetiology

Type I NSTI, also known as polymicrobial NSTI, is the most common NSTI found among patients. In the INFECT study, polymicrobial NSTI was found in 50% of the patients **(M. B. Madsen, Skrede, et al. 2019)**. Polymicrobial infections are generally found in the head, neck, abdomen, and ano-genital regions **(M. B. Madsen, Skrede,**

et al. 2019). Microbes commonly found in polymicrobial infections include *S. aureus*, Viridans streptococci, *S. dysgalactiae*, *Enterococcus* spp., *Escherichia coli*, *Bacteroides* spp., *Prevotella* spp., and *Fusobacterium* spp. **(Skrede et al. 2020)**. Type II NSTI, also kown as monomicrobial NSTI, are primarily caused by *S. pyogenes* or group A streptoccus (GAS). GAS infections accounted for 31% of the patients in the INFECT cohort **(M. B. Madsen, Skrede, et al. 2019)**. Active surveillance of GAS infection already exists in several countries **(G. E. Nelson et al. 2016; Plainvert et al. 2012; Vlaminckx et al. 2004)** and the incidence rate of GAS infections has been reported to be around 3-4 per 100000 person-years **(Steer et al. 2009)**. This incidence rate may be many times higher in certain regions **(Steer et al. 2009)**. Monomicrobial infections are also seen with *S. dysgalactiae* as the causal microbe. *S. dysgalactiae* is a β-hemolytic streptococcus that displays genetic smilarity to GAS **(Skrede et al. 2020)**. In this thesis, the focus has been primarily on type I and type II NSTIs. Type III and type IV NSTI classifications have been suggested, however are not a subject of discussion in this thesis. The causal microbes in type III NSTI include *V. vulnificus*, *A. hydrophila*, and *Shewanella* spp. **(Skrede et al. 2020)**. The causal fungus in type IV NSTI is either *Candida* spp. or *Zygomycetes* **(D. L. Stevens and Bryant 2017; Sartelli et al. 2018; Morgan 2010)**. However, the classification of type IV NSTI is not widely accepted **(D. L. Stevens and Bryant 2017)** Despite these classifications, there is substantial heterogeneity in reported microbial aetiologies and a lack of international consensus on the definitions and pathogenicity **(Eron et al. 2003; D. L. Stevens, Bisno, Chambers, Everett, et al. 2005; D. L. Stevens, Bisno, Chambers, E. P. Dellinger, et al. 2014)**.

In the INFECT consortium, microbes were classified into primary and secondary pathogens **(Skrede et al. 2020)**. Primary microbes were described as microbes that may cause NSTI in patients without the presence of known risk factors whereas secondary pathogens were described as pathogens that may cause NSTI in patients with risk factors **(Skrede et al. 2020)**. Primary microbes included microbes like *S. pyogenes*, *V. vulnificus*, & *Clostridium perfringens*. Examples of secondary microbes included group C and group G streptococci, *S. dysgalactiae*, *Streptococcus agalactiae*, *Streptococcus pneumoniae*, *S. aureus*, *Haemophilus influenzae*, *Neisseria meningitidis*, *Enterobacteriaceae*, *Pseudomonas* spp., *Acinetobacter* spp., *Bacteroides* spp., *Prevotella* spp., *Fusobacterium* spp. **(Skrede et al. 2020)**.

1.2 The human immune response

The microbial aetiology and immune evasion methods evolved by the microbes are only part of the NSTI puzzle. The human immune system's response has also been shown to play a role in NSTI **(Siemens, Snäll, et al. 2020)**. The human immune system is a complex system consisting cells, tissues, organs, functions and interactions between them that has evolved to protect the individual from the environment that co-hosts a universe of pathogenic microbes that are themselves constantly evolving **(Chaplin 2010)**. The human immune system can be split into two functional arms, the innate and adaptive immune systems. The innate and adaptive immune systems represent distinct yet synergistic components of the host's defensive machinery **(Chaplin 2010)**. The innate immune system, encoded by germ-line genes, encompasses physical barriers (like epithelial cell layers that express tight cell-cell con-

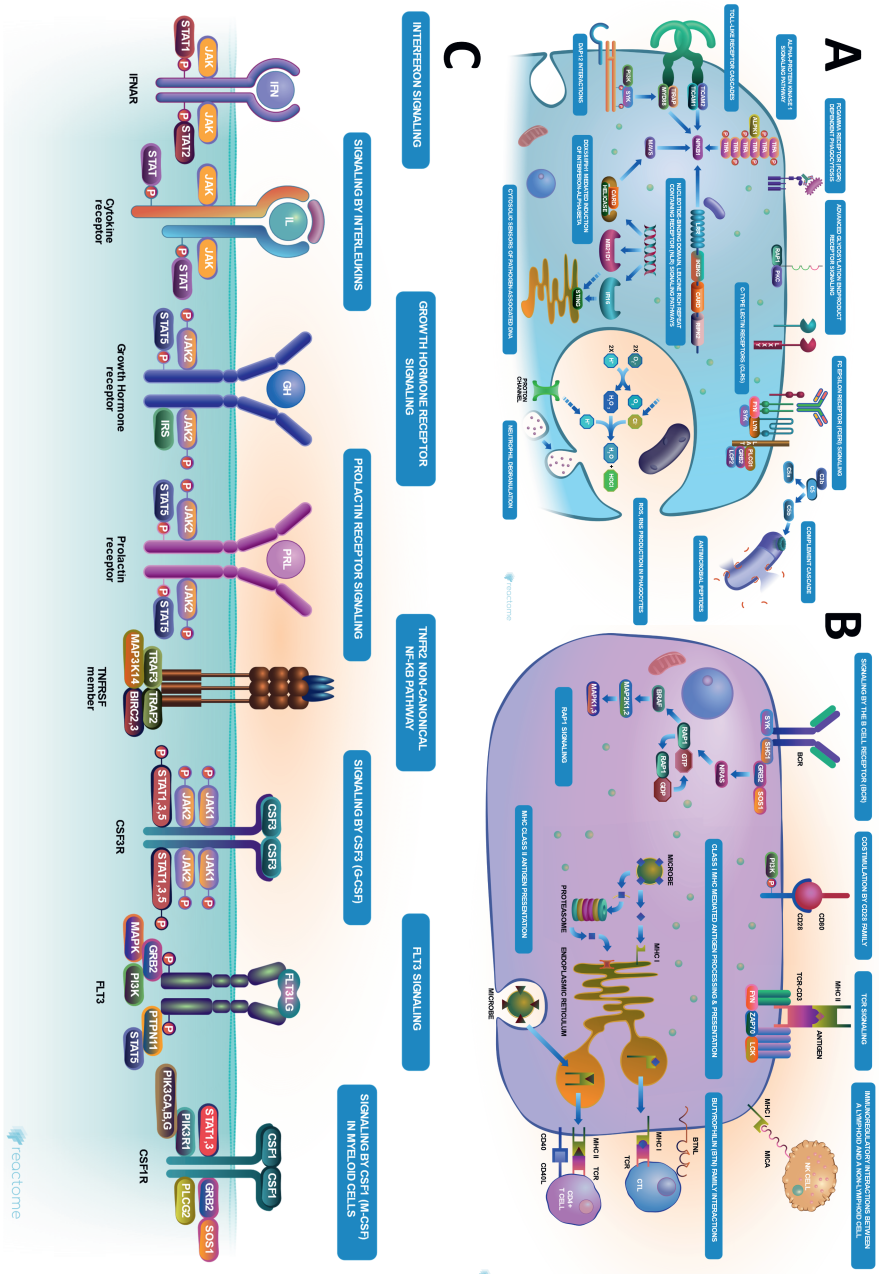


Figure 1.4: An overview of the processes involved in the different arms of the immune system and cytokine signalling from the reactome database **(Sidiropoulos et al. 2017)**. (**A**) is the overview of the processes involved in the innate immune system. (**B**) is the overview of the processes involved in the adaptive immune system. (**C**) is an overview of the cytokine signalling processing. Figure modified from reactome images **(Sidiropoulos et al. 2017)** under the Creative Commons Attribution 4.0 International (CC BY 4.0) License

tacts, secreted mucus layer that overlays the epithelium, and the epithilial cilia that sweep away the mucus layer), bioactive molecules (complement proteins, cytokines, chemokines, lipid mediators, and others), and receptor proteins that recognise microbial structures **(Chaplin 2010)**. This system is constantly active and acts as the host's initial defence line **(Chaplin 2010)**. On the other hand, the adaptive immune system displays remarkable antigen-specificity, leveraging antigen-specific receptors on T- and B-lymphocytes **(Chaplin 2010; Bonilla et al. 2010)**. These receptors are formed by somatic rearrangement of germ-line gene elements, resulting in diverse antigen recognition capabilities **(Chaplin 2010; Bonilla et al. 2010)**. Although the adaptive response becomes more prominent several days post invasion, it relies on the innate system for activation and recruits innate effector mechanisms for robust microbial control **(Chaplin 2010; Bonilla et al. 2010)**. An overview of the human immune system is reproduced from the Reactome database in Figure 1.4 **(Sidiropoulos et al. 2017)**.

The immune response involves numerous leukocyte subsets, originating from hematopoietic stem cells, which differentiate into myeloid stem cells or common lymphoid progenitors **(Huston 1997; Chaplin 2010)**. These then form various immune cells including B cells, T cells, and natural killer (NK) cells **(Chaplin 2010)**. B and T cells are characterised by their specific antigen receptors, with B cells responding to and producing antibodies for varying antigens, whereas T cells binding to processed antigens presented by antigen-presenting cells (APCs) **(X. Chen et al. 2008; Seder et al. 1992; Chaplin 2010)**. Myeloid stem cells give rise to granulocytes, megakaryocytes, platelets and erythrocytes. Granulocytes like neutrophils and macrophages play significant roles in pathogen clearance, inflammation, tissue repair, and are involved in immune regulation through the production of cytokines **(Kennedy et al. 2009; Chaplin 2010)**. Platelets also release immunoligically significant mediators that increases their role in the immune system beyond hemostasis **(Chaplin 2010)**. Other immune cells, including eosinophils, basophils, and mast cells, are involved in allergic responses, immediate hypersensitivity reactions, and host responses to parasites and bacterial infections **(Minai-Fleminger et al. 2009; Schroeder 2009)**. Moreover, cells of the monocyte/macrophage lineage, including dendritic cells, play key roles in the adaptive immune response by processing and presenting microbial antigens, activating T-cell responses, and contributing to antiviral host defense and autoimmunity **(Gilliet et al. 2008; Chaplin 2010)**.

Neutrophils and macrophages play a significant role in the hyper-inflammation and tissue destruction in NSTI patients **(Herwald et al. 2004; Johansson, Linnér, et al. 2009; Kahn et al. 2008)**. T cells and dendritic cells are also often recruited to the site of infections **(Norrby-Teglund, Thulin, et al. 2001)**. This effect has most notably been observed in GAS infections **(Siemens, Snäll, et al. 2020)**. Neutrophils employ several mechanisms at the site of infections including phagocytosis, degranulation, and the formation of Neutrophil Extracellular Traps (NETs) to combat the invading pathogens **(Kolaczkowska et al. 2013)**. Tissue damage occurs as GAS can stimulate neutrophils to release damaging granule effector molecules inducing an apoptotic program along with triggering degranulation **(Borregaard et al. 2007; Kobayashi et al. 2003; Snäll et al. 2016; Soehnlein et al. 2008)**. Streptococcal M protein is one of the major triggers of neutrophil activation and release of granule proteins further downstream **(Gautam et al. 2001; Herwald et al. 2004)**. GAS have been shown to

produce pro-inflammatory cytokines (TNF, IL-6) via the activation of NF-$\kappa\beta$ and MAPK in a MyD88-dependent manner **(Gratz et al. 2008)**. Macrophages (infiltrating highly infected tissue) are proficient in recognising GAS via surface pattern recognition receptors (TLRs) **(Johansson and Norrby-Teglund 2012; Johansson, Thulin, et al. 2010; Johansson, Snäll, et al. 2014; Siemens, Chakrakodi, et al. 2016; Thulin et al. 2006; Valderrama et al. 2018)**. Despite the expectation from macrophages to kill bacteria, the survival, replication, and egress of GAS within macrophages through specific mechanisms such as the manipulation of phagolysosomal degradation further exacerbates the infection and inflammation **(Thulin et al. 2006; Hertzén et al. 2010)**. GAS also trigger both classical (Th1) and alternative (Th2) responses in macrophages, further contributing to the inflammatory state **(Goldmann et al. 2007)**. The systemic inflammation is further exacerbated by superantigens (SAgs) that provoke a massive cytokine storm (production of TNF, IFN-γ, IL-1, IL-2, IL-6, CXCL-8/IL-8, CCL-2/MCP-1, and CCL-3), leading to severe systemic inflammatory responses like Streptococcal Toxic Shock Syndrome (STSS) **(Chatila et al. 1993; Johansson, Thulin, et al. 2010; Emgaard et al. 2019)**.

1.3 INFECT, PerMIT & PerAID projects

The INFECT, PerMIT and PerAID projects are referenced throughout this thesis. The INFECT project was an EU-FP7-Health-funded project between 2013 and 2018. It's goal was to advance our understanding of NSTI using a systems medicine approach. More information can be found at https://permedinfect.com/projects/infect/.

The PerMIT project is an ERAPerMed funded project with a focus on personalised medicine in infections. More information can be found at https://permedinfect.com/projects/permit/

The PerAID project is funded by the Nordforsk initiative and focuses on personalised medicine in acute infectious diseases. More information can be found at https://permedinfect.com/projects/peraid/

1.4 Multi-disciplinary Systems medicine

Systems medicine, akin to personalised medicine, precision medicine, and P4 medicine (P4 referring to predictive, preventive, personalised, and participatory) is an emergent terminology reflecting the escalating inclination towards more systematic, precise, and personalised approaches in medical science **(Hood et al. 2011; Apweiler et al. 2018)**. These methodologies converge on shared objectives: enhancement of diagnostic accuracy, personalisation of therapeutic interventions, prognosis improvement, and more effective prevention strategies. The fulfilment of these goals is facilitated by the comprehensive integration of multifaceted data sources, including individual patient data, clinicopathological variables, and multi-omics datasets **(Apweiler et al. 2018)**.

Systems medicine could be delineated as the application of systems biology for the prevention, understanding, modulation, and recovery from developmental anomalies and pathological events in human health. This discipline emphasises that the

relevance of the models is translationally aimed at diagnostic, predictive, and therapeutic purposes, and requires evaluation on both a medical and biological scale **(Clermont et al. 2009)**. Unlike systems biology which focuses predominantly on the molecular level, systems medicine necessitates the incorporation of mesoscale clinical information, such as clinical variables, and biomarkers into its models. It aims to bridge diverse organisational levels from molecules to populations, incorporating key factors into translational models **(Clermont et al. 2009)**. Despite not originating from clinical medicine, systems medicine draws its relevance from it and also benefits from systems biology advancements **(Clermont et al. 2009)**.

The integration of clinical information in systems biology and systems medicine has the potential to challenge the "one size fits all" philosophy, potentially supporting the development of more personalised therapy designs and improving the risk to benefit ratio. One could argue that medicine has never targeted a one size fits all model and that medicine has always been "personalised" in its approaches with the diagnoses, prognoses and therapies focusing on the individual and their specific conditions **(Apweiler et al. 2018)**. However, the comprehensive understanding a systems medicine approach offers **(McCarthy 2004; Wolkenhauer et al. 2013; Korcsmaros et al. 2017)** via the integration of disparate data sources (from patient records to sequencing and multi-omics data) and external data bases/sources encompassing aspects of molecular pathways, gene-regulatory elements, or drug targets promises substantial differences from the traditional approaches in medicine **(Elefsinioti et al. 2016; Ghosh, Matsuoka, et al. 2011; Gomez-Cabrero et al. 2014)**. Systems medicine approaches often strive to fully exploit the IT infrastructure set up to manage data storage, provenance, security, and data sharing in order to bridge the gap between computational research and medical research and ultimately between research and patient care **(Apweiler et al. 2018)**. This data amalgamation and its interpretations are achieved through integrative workflows incorporating various statistical, mathematical, computational, machine learning, and data science techniques. Due to the application being of a medical nature, it is also imperative that these workflows maintain the strictest standards statistically **(Apweiler et al. 2018)**.

1.5 Computational methods: from systems biology to systems medicine

1.5.1 Systems biology

Systems biology is an interdisciplinary research field aimed at understanding the intricate interplay among biological components, which range from molecules to entire species. This field leverages quantitative measurements, genomics, bioinformatics, proteomics, and mathematical models to predict the dynamic behaviour of biological systems **(Bruggeman et al. 2007)**.

Advancements in omics technologies have fostered two primary, and somewhat contrasting, approaches in systems biology: the top-down (deductive) and bottom-up (inductive) methodologies **(Saccenti and Svensson 2020)**. The top-down approach capitalises on the abundance of system-wide data generated by these advanced technologies, extracting valuable insights through the use of statistical, machine learning

methods, and network analysis. This enables researchers to understand the intricate interplay between various molecular constituents such as genes, proteins, and metabolites, thus illuminating the overarching behaviour of the biological system across diverse conditions **(Saccenti and Svensson 2020; V. A. Martins dos Santos et al. 2020)**. Conversely, the bottom-up approach is rooted in the deep-seated molecular and biochemical knowledge of specific biological mechanisms. Leveraging this knowledge, researchers can create mathematical models that reproduce experimental data, thereby predicting the system-wide behaviour. Both approaches are integral to systems biology and the data-integration-modelling-validation cycle introduced by **(Rosato, Tenori, Cascante, De Atauri Carulla, et al. 2018)**.

The transition from data-poor (patient data) to data-rich (molecular characterisation) applications, facilitated by the advent of high-throughput genomics, transcriptomics, metabolomics, and other "omics" disciplines, has enabled Systems Medicine approaches **(Noble 2008)**. This influx of data, however, presents both an opportunity and a challenge due to the heterogeneity observed in both the data and patients **(Saccenti and Svensson 2020; V. A. Martins dos Santos et al. 2020)**.

Systems Medicine offers a different approach, emphasising reciprocal feedback between clinical investigations and multiscale computational, statistical, and mathematical analyses **(Saccenti and Svensson 2020; V. A. Martins dos Santos et al. 2020)**. This approach complements the traditional reductionist paradigm and has been tauted for improvements in the identification of disease-related mechanisms, novel drug targets and biomarkers, and improved patient risk assessment **(Saccenti and Svensson 2020; V. A. Martins dos Santos et al. 2020)**.

1.5.2 Statistics

The disciplines of systems biology and systems medicine fundamentally depend on advanced statistical techniques for deciphering the complexity of biological and pathological systems. These techniques enable the interpretation of extensive, high-dimensional data and provide a means to derive significant patterns from big data sources. The ability to uncover various relationships between multiple variables is crucial, as is the application of these statistical methods across data management and analytical pipelines. Such pipelines encompass tasks like sample size determination, addressing imputation issues, standardising and normalising procedures, data summarising techniques, conducting hypothesis testing, and determining the statistical significance of results **(Yan et al. 2017)**. Furthermore, statistical and probabilistic methods serve an essential role in counteracting unintentional biases and confounding effects that may occur between the experimental and control groups **(Sullivan et al. 2016)**. The use of statistical and probability-based designs and distributions formulates the assumptions underpinning variables and hypothesis testing and establishes the stringency criteria **(Viti et al. 2015)**.

1.5.3 Machine learning

The concept of machine learning was initially introduced by Samuel, who sought to enable a computer to play checkers autonomously **(Samuel 1959)**. Despite the constraints of computational power during that period, he introduced pioneering al-

gorithms resembling modern tree-based and curve-fitting approaches to facilitate automated decision-making processes **(Samuel 1959)**. Machine learning models, driven primarily by data **(Rosmalen et al. 2022)**, share a somewhat blurry boundary with statistical models. However, they fundamentally differ in their objectives. Machine learning models primarily aim to automate the development of predictive models from data such that the models can be evaluated objectively on new data. While these models can integrate domain-specific expert knowledge **(Vidulin et al. 2014)**, the primary emphasis lies on data-driven automation capable of discerning complex patterns that may otherwise be difficult to identify **(Rijn 2016)**.

Machine learning algorithms are typically categorised into two main types: supervised and unsupervised learning algorithms. Supervised learning algorithms leverage known inputs and outputs during training, enabling them to make predictions on new, unseen datasets **(Leist et al. 2022)**. Conversely, unsupervised learning algorithms are designed to unearth hidden patterns and intrinsic structures within the data **(Leist et al. 2022)**. Supervised learning algorithms can further be delineated as regression or classification algorithms, contingent upon whether the predicted variable is continuous or discrete in nature **(A. Singh et al. 2016)**. Over the years, an assortment of unsupervised methods has been cultivated for hierarchical clustering and dimensionality reduction purposes **(An et al. 2023)**. Such methods include, but are not limited to, K-means clustering **(Hartigan et al. 1979)**, Euclidean distance **(Murtagh et al. 2012)**, DBScan **(Birant et al. 2007)**, Principal Component Analysis (PCA) **(Maćkiewicz et al. 1993)**, and t-Distributed Stochastic Neighbor Embedding (t-SNE) **(Van der Maaten et al. 2008)**. In recent decades, supervised learning strategies have burgeoned, introducing a wealth of diverse ideas, metaphors, and interpretations. Nevertheless, two fundamental strategies have emerged as particularly effective: (1) the mapping of variables and measurements within a multi-dimensional Cartesian space and subsequently discerning curves to separate groups or fit the measurements, and (2) conducting a sequence of recursive binary splits on the data to construct predictions based on the terminal splits.

Support Vector Machines (SVMs) exemplify the former approach, as this method seeks to delineate a hyperplane in a multi-dimensional space, segregating classification groups while maximising the margin between the hyperplane and the nearest members of each group **(Vapnik et al. 1996; Schölkopf et al. 2002)**. Numerous methods also endeavour to fit curves that approximate the original data distribution. Regression models, both linear **(Yu et al. 2017)** and logistic **(LaValley 2008)**, are representative of such approaches, but neural networks add a significant amount of flexibility in this strategy by relinquishing the assumption of a linear relationship between inputs and outputs **(Murphy 2022)**. Neural networks were originally designed to mimic the behaviour of biological neurons **(Fukushima 1975)**. Fundamentally, the method begins with an activation function and allow a series of multiplications and additions to adjust the function to fit the data. Optimization techniques such as backpropagation **(Linnainmaa 1976; Leung et al. 1991)** are then employed to refine the parameters (multipliers and additions) for a given training dataset. Various versions of this method with several different names under the banners of neural networks and deep neural networks have been devised based on the difference in activation functions **(Rasamoelina et al. 2020)** (RelU **(Agarap 2018)**, ELU **(Clevert et al. 2015)**, GELU **(Hendrycks et al. 2016)**, Swish **(Ramachandran et al. 2017)**, Sigmoid **(J. Pen-**

nington et al. 2017), Leaky-RelU **(K. He et al. 2015)**, etc.), the number of layers **(Samek et al. 2021)**, the characteristics of output nodes **(Amato et al. 2013)**, and the type of data that the model handles (tabular data **(Murphy 2022)**, images **(J. Gu et al. 2018)**, sequences **(Medsker et al. 2001; Vaswani et al. 2017)**, graphs **(J. Zhou et al. 2020)**, and so forth).

Classification and Regression Trees (CART) **(W.-Y. Loh 2011)**, often referred to as decision trees, exemplify the latter approach **(A. J. Myles et al. 2004)**. This methodology operates by recursively partitioning the input data space and subsequently defining a localised model within each resulting subset **(Fellinghauer et al. 2013)**. These models are frequently visualised as a metaphorical tree, with splits and terminal ends denoted as branches and leaves within the scope of the metaphor. Beyond their precision, these methods have attracted interest due to their high interpretability **(Kazemitabar et al. 2017)**. The fusion of this procedure with other machine learning principles in an ensemble approach has led to the development of algorithms with exceptional performance. The Random Forests algorithm **(Breiman 2001)**, for instance, is predicated on the assumption that applying the same learning procedure to various data subsets yields sufficiently diverse models. By embracing the concept of "wisdom of crowds **(Fleenor 2006)**" this algorithm combines bootstrapping and an ensemble of decision trees **(Breiman 2001)**. Random Forests has evolved into a robust, highly successful and frequently used algorithm with the possibility of even calculating the significance of the interpretable results **(Archer et al. 2016)**. Certain methods such as AdaBoost strive to accelerate the process speed embodied in random forests by restricting the depth of individual decision trees to two leaves (termed "stumps") and modulating the influence that each decision tree exerts **(Freund et al. 1997)**. Similarly, approaches like gradient boosting construct decision trees based on the coefficients of a loss function as opposed to the data itself **(Natekin et al. 2013)**. In 2015, Chen et al. introduced a technique named XGBoost (eXtreme Gradient Boosting), which amalgamated several selling features from the aforementioned algorithms with out-of-core computation methods and greedy optimisation techniques **(T. Chen, T. He, et al. 2015; T. Chen and Guestrin 2016)**. The combination of these elements has led to XGBoost emerging as the algorithm of choice for many winning projects in machine learning competitions as XGBoost surpassess the speed of AdaBoost, while maintaining the accuracy of random forests **(T. Chen, T. He, et al. 2015; T. Chen and Guestrin 2016)**.

1.5.4 Modelling approaches: Networks

Computational models, designed to symbolically encapsulate (biological) entities and systems, serve as instrumental tools for comprehension, analysis, and prediction of the behaviours of these entities and systems. Classifying and understanding computational modelling approaches can be achieved from multiple perspectives. These perspectives can include the point of view of particular concept (eg. probability theory **(Murphy 2022)**), the reliance of the model on a priori knowledge or empirical data **(Rosmalen et al. 2022)**, or the quantity of independent variables that instigate changes within the modelled system use. In this section, I introduce some modelling concepts from the prospective of network theory. However, detail descriptions of the exact methodology used and the rationale behind their choices can be found in the

methods sections of their respective chapters.

Networks offer a comprehensive strategy for modelling intricate and expansive systems through their constituent entities and interconnections **(Jahagirdar, Suarez-Diez, et al. 2019)**. Essentially, a network is a graph where nodes signify the constituent entities and edges embody the relationships, associations, or interactions among them. Network models are capable of portraying a variety of interactions, from the simplest binary relationships **(Wille et al. 2006)** to both linear and non-linear dynamics in relation to time and space, encapsulated within parameters of ordinary differential equations (ODEs) **(Jahagirdar, Suarez-Diez, et al. 2019)** and partial differential equations (PDEs) **(Göttlich et al. 2005)**, from statistical associations **(Friedman et al. 2012)** to complex multi-faceted interactions **(Ray et al. 2016)**. The approach of systems medicine has enthusiastically adopted this style of network modeling, capitalizing on its capacity to encapsulate behaviors of complex systems constituted by heterogeneous variables that represent a multitude of measurement types **(Gustafsson et al. 2014; Auffray et al. 2009)**.

As the adoption of network models has become increasingly prevalent, the field of data-driven network inference, especially based on statistical associations or interactions, has seen considerable advancement. Although sophisticated inference algorithms have proliferated, a large number of these algorithms still rely on either covariance-based computations or information theory-based probabilities to evaluate the associations within these algorithms **(Marbach et al. 2012)**. For a detailed introduction on network inference methods, I would guide the reader to another study where we review and benchmark multiple statistics and machine learning based network inference methods against data generated from a dynamic model of arachidonic acid metabolism **(Jahagirdar, Suarez-Diez, et al. 2019)**. Each chapter in this thesis addresses interpretations of network models in accordance with the methods employed to infer the networks. For a succinct overview of biological network interpretation, readers are encouraged to refer to Figure 1 in **(Jinawath et al. 2016)**.

1.6 ReadMe.md

This thesis sits at the precipice of a multidisciplinary intersection, encompassing (a) Computational systems biology, which includes methods from data science, machine learning, statistics, and systems medicine; (b) Medical research and practice, with a focus on infectious diseases, emergency medicine, and medical surgery; and (c) (Micro-)biology, emphasising pathogenesis, immunology, and host-pathogen interactions. This thesis strives to enhance the integration of computational systems medicine approaches, particularly modelling, data science, and machine learning methods with the established medical and microbiological research practices used in the understanding and management of NSTI.

The aim of this thesis is to formulate a structured scientific enquiry using computational approaches to understand and explore the underlying biological mechanisms in necrotising soft tissue infections. I advance this aim by

- Advancing the use of computationally derived modelling, data science and machine learning methods in the biomedical research surrounding NSTI.

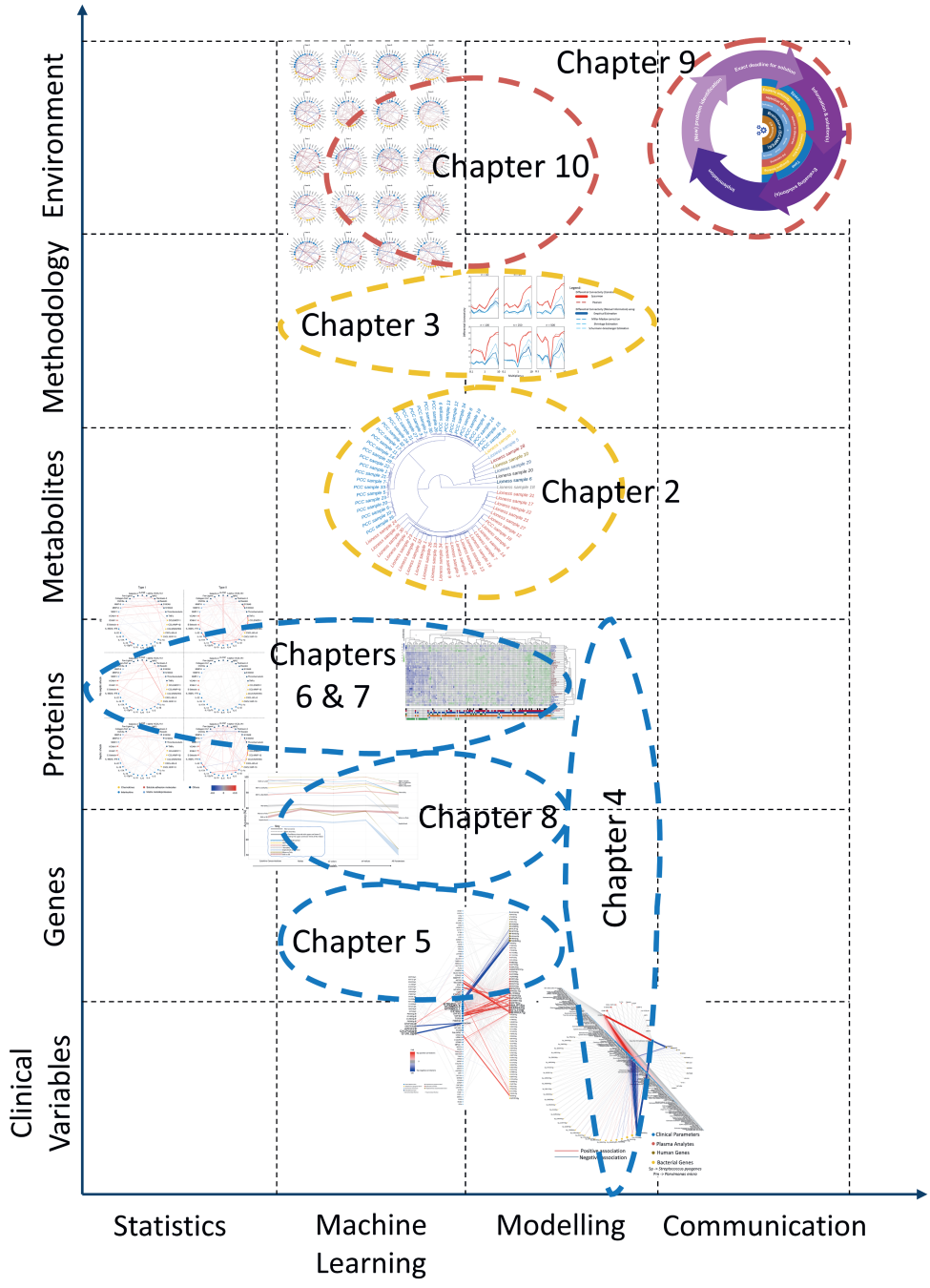


Figure 1.5: A pseudo-graph conceptualising the chapters of this thesis based on the elements that were studied in the chapter and the methodologies used to study those elements. The X-axis (horizontal) on the graph represents the methods used in the chapter and the Y-axis (vertical) represents the elements studied in the chapter. From the graph, it can be interpreted that in **Chapter 4**, I studied clinical variables, genes, and protein measurements of NSTI patients using modelling approaches.

- Using said computational approaches to investigate data from NSTI patients to promote data-driven hypotheses exploring the underlying mechanisms of NSTI.
- Employing computational approaches for mining heterogeneous multi-omic big data sets, with the objective of identifying variables that exhibit predictive and discriminatory value between different NSTI sub-types, aetiologies, and patient outcomes.

In **Chapter 1** (*Introduction: Why this matters?*), I start with a brief introduction to the philosophical discussions that have taken place while defining diseases. I further introduce necrotising soft tissue infections (NSTI) and guide the reader through the research concepts that are important to understand the relevance of this thesis. I end **Chapter 1** by providing a useful glossary of terms from different disciplines to aid in the reading of this thesis. The concept of network modelling and its role in systems medicine is introduced in **Chapter 1**, however, it is immediately clear that in constructing these networks, the subject-specific information is lost. In **Chapter 2** (*Personalising Metabolomics? A Closer Look at Single Sample Network Inference*), we explore, review and benchmark methods that claim to recover these individual-specific networks in the context of metabolomics. Although, these concepts were developed using one specific network inference method, we realise that the question of inferring networks can be separated from the core concept of calculating the individual-specific networks. In **Chapter 1**, a succinct overview of network inference methodologies is presented, elucidating that despite the complexity of inference algorithms, a significant majority rely fundamentally on correlation and mutual information to ascertain associations between variables. **Chapter 3** (*Correlation or Mutual Information? That is the Question!*) delves into an evaluation of the capacity of these two association measures to identify significantly differentially associated metabolites. This computational experiment was iteratively conducted on various metabolomic datasets, data derived from a dynamic model, and data created through established correlation structures. This evaluation plays a role in the selection of association measures downstream in this thesis. Figure 1.5 visualises all the chapters based on the elements that were studied and the methods used to study them.

Chapter 4 (*Eventually it Boils Down to the Clinical Variables*) tackles the problems presented by mixed data types and low sample sizes found in NSTI data sets to analyse relationships between variables measured in the clinic and intensive care unit (ICU), gene transcriptimic measurements from humans and bacteria, and proteins measured in the plasma. **Chapter 4** focuses more on the associations with clinical variables as those variables are readily accessible and can help enable timely diagnosis and treatment. Several important host and pathogen genes are exposed for having strong associations with clinical variables in **Chapter 4**. An in-depth examination of these genes is conducted in **Chapter 5** (*Inside the Inferno: An Inspection of the Host-Pathogen Interactions*), wherein an interactome is constructed using dual RNA-Seq gene transcriptomic profiles derived from biopsy samples of NSTI patients. We find aetiology-dependent responses and were able to postulate modes of entry and immune evasion strategies of *S. pyogenes*. We also find associations with genes coding for cytokines.

An exhaustive analysis of these cytokines is conducted in **Chapter 6** (*The Race*

Against Time: Discriminatory Plasma Biomarkers) and **Chapter 7** (*The Immune System Responds: Systemic Immune Activation Profiles*), employing a synergistic approach of statistical, machine learning, and network-based methods. This ensemble framework is utilised to scrutinise and contrast the levels of various chemokines, interleukins, soluble adhesion molecules, matrix metalloproteases, growth factors, and other small proteins measured in plasma against various controls. **Chapter 6** is centred around distinguishing between NSTI and non-NSTI cases, whereas **Chapter 7** prioritises differentiating between NSTI and Cellulitis. The two chapters also address differentiating between the occurrence of septic shock and between various microbial aetiologies. In these chapters we identify discriminatory plasma biomarkers with potential value for diagnostic, prognostic, and therapeutic approaches in NSTI. **Chapter 8** (*Deadly Dance: Understanding the Interplay of Pro- and Anti-Inflammatory Cytokines*) presents an exploratory discourse aimed at investigating whether the pairwise relationships among the measured cytokines carry significant insights pertaining to our mechanistic understanding of bacteremia in NSTI. Ratios and model parameters are employed in **Chapter 8** as proxies to represent a diverse range of interactions between the cytokines, incorporating the dual nature of cytokines as both pro-inflammatory and anti-inflammatory agents. These pair-wise relationships are used to discriminate between NSTI sub-types, microbial aetiologies, and patient outcomes.

Chapter 9 (*Wisdom of the Informed Crowds*) draws upon (a) the collective insights gained from the work conducted in the preceding chapters, and (b) the considerations of scholarly literature devoted to multi-disciplinary problem-solving being focused solely on managerial aspects. **Chapter 9** introduces a problem-solving model tailored for multi-/inter-/trans-disciplinary research addressing technical aspects while refraining from the reliance on heuristic methods of individual fields. **Chapter 9** also delves into straightforward strategies borrowed from the realm of software engineering, aiming to steadily mitigate the mathematical linguistic barrier in the long run. A key component of the proposed model underscores the utilisation and necessity of human-interpretable, data-driven visualisations, extending beyond mere communication to fortify reliable decision-making processes. Based on this tenet, I was awarded an internal DS/AI fellowship/grant, enabling the research carried out in **Chapter 10** (*Automated NETwork VISualisation with anvis*). In **Chapter 10**, I develop a software package (anvis) that generates multiple human-readable coordinate visual representations of networks. To achieve this without necessitating manual curation, I leverage data science and machine learning methods, thereby rendering the software adaptable to any user-defined data analysis pipeline. **Chapter 11** (*Discussion: Piecing it Together*) offers a recap of the reasoning that underpins each research component within this thesis, coupled with a critical examination of the results and their implications. **Chapter 11** also delves into the consequential considerations and ethical implications associated with deploying artificial intelligence methods within this field. The discourse concludes by contemplating both the limitations and strengths of the research undertaken within this thesis, elaborating on how these aspects influence the findings presented herein.

Lexicon Library

This thesis sits at the precipice of a multidisciplinary intersection, encompassing (a) Computational systems biology, which includes methods from data science, machine learning, statistics, and systems medicine; (b) Medical research and practice, with a focus on infectious diseases, emergency medicine, and medical surgery; and (c) (Micro-)biology, emphasising pathogenesis, immunology, and host-pathogen interactions. Accordingly, this thesis employs a broad array of specialised terminology and jargon drawn from multiple disciplines. To facilitate seamless comprehension, I provide a comprehensive glossary of potentially unfamiliar terms, aiming to assist readers coming from diverse backgrounds.

Adjacency matrix	An adjacency matrix (also known as connection matrix) representing a network (or a graph) is a matrix with rows and columns labelled by the network nodes/vertices and the values inside the matrix representing the edges/connections between the nodes.
Aetiology	Aetiology (also written as etiology in the US) refers to the cause of the disease. In this thesis, microbal aetiology refers to the identification of the bacteria present in the host responsible for causing the symptoms.

Bias	Bias in statistics (relevant to this thesis) refers to a systematic error that can cause differences between the results of an analysis and the potential true representation of the samples. Bias can be introduced in many forms throughout the entire process of the data analysis including data collection, imputation, estimation, visualisation, and interpretation.
Bivariate	Bivariate analyses in statistics are analyses determining pair-wise relationships between two variables.

Cellulitis	Cellulitis is a bacterial infection of deeper layers of the skin and the underlying tissue. Cellulitis can be life-threatening if untreated, however, mortality rates are not comparable to NSTIs. Cellulitis is relevant in this thesis because the symptoms presented during cellulitis are similar to those presented during NSTI.
Comorbodity	Comorbidity is the presence of one or more conditions present in the patient in addition to the primary condition of interest. As an example in this thesis, for a patient suffering from NSTI, the additional presence of type II diabetes is a comorbidity.

Conditional dependence/ independence	In probability, conditional dependence is the presence of a relationship between two variables when the relationship of all the other variables on those initial variables is discounted for. Similarly, if variable X does not affect Y in the presence of other variables Z, then X and Y are said to be conditionally independant.
Connectivity	In networks, connectivity is a measure accounting for the number and strength of connections that a particular node makes in the network. The connectivity value is a property of the node in the network.
Correlation	Correlation (Pearson's correlation) is a linear measure of association between two variables. This is a normalised measure of co-variance and the values can vary between [1,-1]. A 0 value represents no association.
CXCL-10/IP-10	CXCL-10/IP-10 is a chemokine (cytokine) involved in the chemotaxis of various cells. It is secreted by monocytes, endothelial cells and fibroblasts in response to IFN-γ. In this thesis, CXCL-10/IP-10 was recognised as a discriminatory biomarker between Type I and Type II NSTI and is found important in several analyses associated with NSTI types
Cytokines	Cytokines are a broad and loose category of small proteins that include chemokines, interferons, interleukins, lymphokines, and tumor necrosis factors. These proteins are involved in cell signalling via cell signalling receptors. Cytokines play an important role in the host immune response, infection, inflammation, trauma, and sepsis.
D	
Decision tree	In machine learning, a decision tree is a non-parametric supervised learning algorithm utilised for classification and regression tasks. The method is based on performing a sequence of repetitive binary partitions of a data set and then inferring predictions from the ends of the terminal split. This structure is metaphorically visualised as a mirrored tree and the splits and decision points are metaphorically referred to as leaves and branches. The decision tree also forms an independent fundamental unit in many complex machine learning algorithms.

Dummy parameters

Dummy parameters in computer science are parameters representing discrete data measurements/observations by converting it into values of 1 and 0. In this thesis, we use the one-hot encoding method to create dummy variables. In this method, we create a feature for every factor level in the discrete data type. These features contain information regarding the presence (represented by 1) or absence (represented by 0) of that particular level for each sample.

ELISA

In biochemistry, enzyme linked immunosorbent assay is commonly used to detect the presence of a protein in a liquid sample using antibodies directed against the protein to be measured. The concentrations of IL-23 and IL-33 used in this thesis were measured using ELISA

FDR

In statistics, the false discovery rate is a method conceptualising the rate of type I errors. This is particularly done when supervising multiple test comparisons.

Gaussian (Normal) distribution

In probability, a Gaussian or normal distribution is a probability distribution that is symmetric around the mean. The number of data points are highest closest to the mean and asymptomatically lower further away from the mean. This type of a distribution has also been described as a bell curve.

GGM

A Gaussian graphical model represents conditional dependencies or partial correlation coefficients between multiple variables in the form of a network where the nodes are normally distributed variables and the non-zero edges represent conditional dependencies.

GO enrichment

Gene Ontology enrichment is a method of interpreting the functions associated with a set of genes based on the pre-defined gene ontology classification system.

GRaFo

GRaFo in machine learning is a method to build conditional independence graphs using the random forest algorithm and a stability selection procedure based on a probabability measurement.

Hierarchical clustering

In data science, hierarchical clustering is a method to build a hierarchy of unsupervised clusters formed based on the properties associated with the samples. Hierarchical clustering can follow both top-down (performing recursive splits) and bottom-up (merging of clusters) approaches.

I

IL-6 IL-6 is an interleukin (cytokine) secreted by macrophages in response to surface recognition molecules recognising the presence of *S. pyogenes*. It acts as a pro-inflammatory cytokine that has been associated with the severity of NSTI.

INFECT Study The INFECT study was the world's largest multicentre, prospective observational study on NSTI patients that was active from 2013 to 2018 and was supported by the EU-FP7-Health framework. Most of the data analysed in this thesis was collected from NSTI patients enrolled in the INFECT study.

Inference In data science, inference methods can be used as a terminology to group together a cluster of different algorithms using various methodologies aimed at calculating and inferring the associations/interactions between different variables. These methods are particularly useful as methodologies to infer complex networks showing interactions between several variables.

Inflammation Inflammation is a normal part of a complex biological response of the body's defence against harmful stimuli and injury. Inflammations can be detected by outward symptoms that involve heat, pain, redness, and swelling. Inflammation helps protect the body by localising and eliminating the injury-causing agents, removing damaged components, and initiating the recovery process. Chronic inflammation can be harmful to the body.

Interactome An interactome is the whole set of interactions that occur within a particular cell, body or system. Interactomes can be built using interactions among genes, proteins or any other biological unit. Mapping and analysing interactomes help us infer complicated networks and relationships within the entire system.

J

Joint probability distribution A joint probability distribution describes the likelihood of different combinations of outcomes for two random variables occurring simultaneously given the two variables are defined in the same probability space. The joint probability distribution does not overrule each variable's independent probability distributions, expected values, variances, and standard deviations but instead provides a useful platform to study relationships between multiple random variables and the dependencies among them.

K

Kruskal-Wallis test — The Kruskall-Wallis test is a nonparametric test to figure out if the samples originated from the same distribution with the null hypothesis stating that the mean ranks of the groups are the same. The Kruskal-Wallis test does not assume a normal distribution of the underlying data.

L

LIONESS — Linear Interpolation to Obtain Network Estimates for Single Samples is a single sample network inference method introduced by Kuijjer et al.

Luminex multiplex assay — Luminex multiplex assay is an immunoassay that precisely measures analytes in one sample. The protein/cytokine/plasma analyte concentrations used in this thesis were measured using this assay.

M

M protein — M protein is an important virulence factor expressed on the surface of *S. pyogenes* and plays multiple roles in streptococcal infection, including resistance to phagocytosis.

Mann-Whitney test — Mann-Whitney U test or Wilcoxon rank-sum test is a nonparametric statistic test that compares two statistic means that come from the same population. The null hypothesis under this test is taken as the two distributions are identical or that the probability of a randomly selected value from the first variable being greater than the 2nd variable is the same as the probability of a randomly selected value from the second variable being greater than the first variable.

MGM — Mixed graphical models are network models representing conditional dependancies for the edges. However, mixed graphical models allow for the incorporation of one set of variables to follow a Gaussian distribution and another set of variables to follow a multinomial distribution.

Monomicrobial — Mono-microbial infections are infections where there is a single causal organism. The NSTI infections caused by a single bacteria are known as Type II NSTI or monomicrobial NSTI. In this thesis, this criterion is used in many chapters to differentiate patients based on microbial aetiology.

Mutual Information	In information theory, mutual information (MI) is a non-linear measure of association between two variables. It is based on the quantification of the amount of information gained about one random variable by observing another random variable. MI association takes the values $[0, \infty]$

N

Network	Network models are a holistic approach to model large complex systems by providing an uncondensed description of the underlying units and the interactions between those units. These units are depicted as nodes and the interactions are depicted as edges in the network. This system can be visualised as a graph in the cartesian space to communicate the system functions.
Network hubs	In a network, hubs are nodes that have significantly more edges/connections compared to the number of connections an average node has in the network.
Network motifs	Network motifs are sub-graphs in a network that represent a topological pattern that can repeat itself multiple times throughout the same network or many other networks. These patterns may be interpreted as frameworks by which a particular function could be achieved, for eg. a positive feed back loop.
NF-$\kappa\beta$ signalling	NF-$\kappa\beta$ is a protein complex that plays a crucial role in cytokine production, cell survival and regulating many aspects of the immune response to an infection.
Non-NSTI	In this thesis, Non-NSTI patients refer to the patients that were initially suspected of having NSTI but later found to be non-necrotic upon surgery. In this thesis, the Non-NSTI patients are used as a control to differentiate from NSTI patients as they showed similar initial symptoms.
NSTI	Necrotising soft tissue infections, also referred to as necrotising fasciitis, are bacterial infections characterized by extensive damage to soft-tissue compartments. Although these infections are rare, they are associated with a high risk of mortality, severe long-term disability, and a loss of quality of life. They are a significant health burden as the initial diagnosis is challenging due to vague symptoms and a lack of specific diagnostic tools. They are often colloquially called "Flesh-eating disease

O

Observational study	An observational study in a medical context is an investigative effort to collect data on health-related participant outcomes without directly assigning specific interventions. They are used to examine predetermined treatments, interventions, and policies.
ODEs	Ordinary differential equations (ODEs) in systems biology are differential equations that are commonly used to model dynamic systems that vary in time, such as the change of concentration of a substance in time. The differential equations have a single independent variable. A dynamic system can be modelled by writing a system of ODEs and optimising the parameters such that the model behaviour matches reality.

P

Pathway enrichment analysis	In computational biology, pathway enrichment analysis is a method that identifies biological pathways/functions that are overrepresented in a gene list and sorts them by relevance.
Patient stratification	Patient stratification is the classification of a patient population into distinct sub-groups on the basis of the presence, difference, or absence of relevant characteristics in a disease. This stratification can help with decisions such a precision treatments for different groups.
PBMCs	Human peripheral blood mononuclear cells are blood cells isolated from healthy donors. Some of the validation data used in this thesis was generated from in-vitro stimulation experiments using PBMCs.
PCA	Principal component analysis in statistics is a dimensional reduction method that is used to summarise the contents of a large data set by creating a set of "summary indices" that can be more easily visualised in two dimensional space.
PCLRC	PCLRC is an algorithm that can be used to add a probabilistic measurement for edges in congruence with an association measure used in any network inference method. The probabilistic measurement of edge likeliness is interpreted as a confidence level based on which we could accept or reject the measured association between two random variables in the network. This chosen threshold can also act as a stringency measure to determine the interactions that could be studied.
Personalised medicine	Personalised medicine or precision medicine are buzzwords used to describe medical models that aim to provide tailor-made treatments and strategies for individuals or small groups of individuals based on their characteristics.

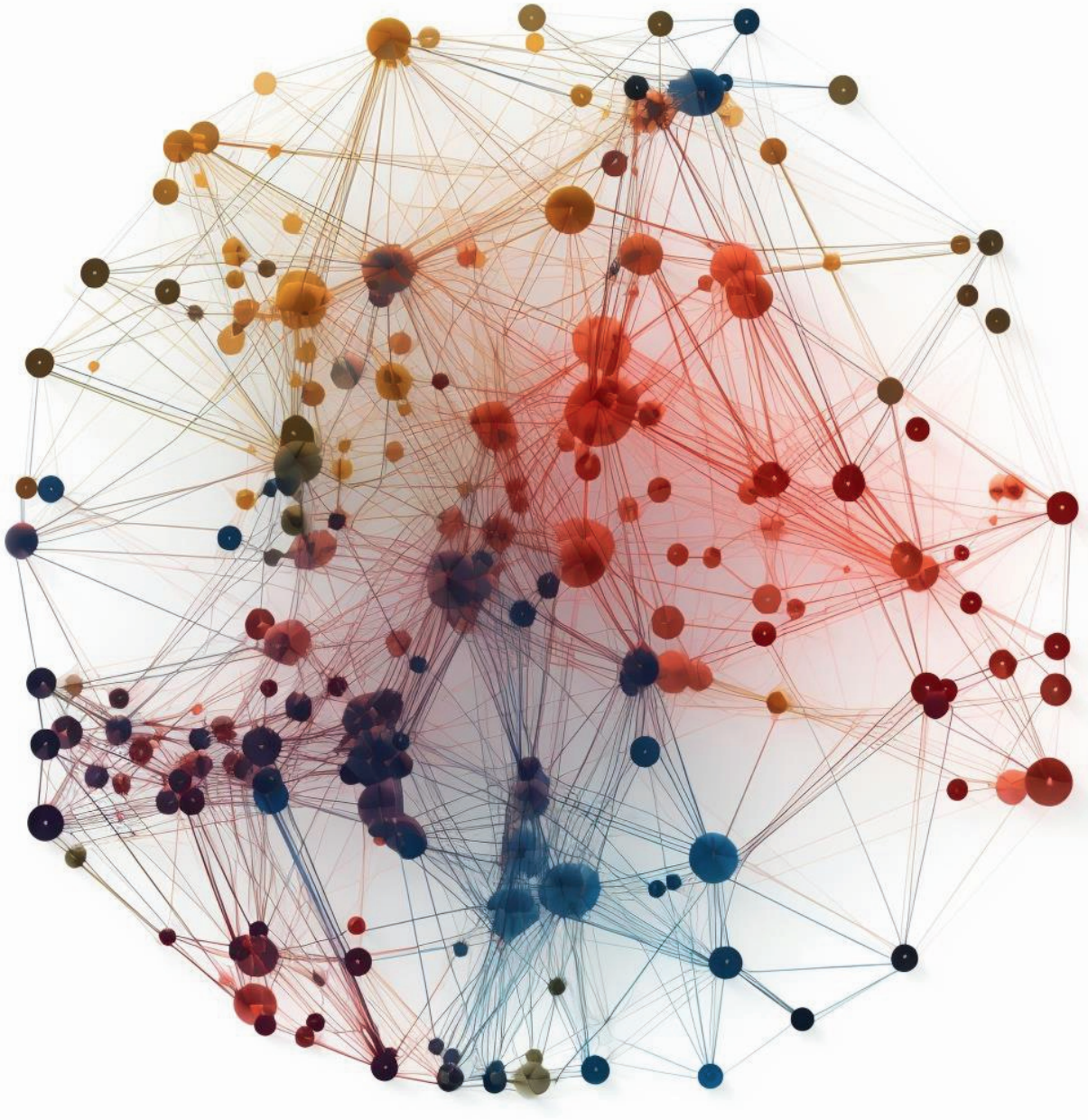
Polymicrobial	Poly-microbial infections are infections where there are many causal organisms. The NSTI infections caused by multiple bacteria are known as Type I NSTI or polymicrobial NSTI. In this thesis, this criterion is used in many chapters to differentiate patients based on microbial aetiology.
Power (statistics)	In statistics, power is the probability of not making a type II error or getting a false negative. In other words, power refers to the probability that a hypothesis test can detect an effect in a sample when the same effect exists in the population.
Pro-inflammatory/ Anti-inflammatory	In relation to cytokines, pro-inflammatory cytokines are small proteins that play a crucial role in modulating inflammation by stimulating, recruiting, and proliferating immune cells. Anti-inflammatory cytokines are small proteins that play a role in the immunoregulation of pro-inflammatory cytokine responses.
Prognosis	In medicine, prognosis is the expected development of the disease. This development includes the rate of worsening (or improvement) of symptoms and quality of life aspects.
Prospective study	In medicine, prospective studies are studies that follow patients over time. Data and measurements are collected with respect to time as their characteristics and conditions evolve.
P-value	In statistics, a p-value is a measurement used in hypothesis testing. It represents the probability of obtaining the observed result assuming the null hypothesis is true. The null hypothesis here is that the distributions of the two data are the same. Hence lower the p-value, the more unlikely the observed results, hence more significant. The p-value makes no claims on the magnitude of the differences, only on the significance.
R	
Random Forests	In machine learning, random forest is an algorithm based on an ensemble of decision trees. The algorithm is based on a resampling approach to build a number of decision trees and counting the number of decision trees supporting each prediction. The algorithm then tallies these votes for an overall prediction.

ROC curve (AUC)	A receiver operating characteristic curve is a graphical plot between sensitivity (also known as true positive rate) and 1-specificity (also known as true negative rate). This plot illustrates the diagnostic ability of a binary classifier as its discrimination threshold is taken from the minimum value to the maximum value. The area under the curve (AUC) of this curve can be interpreted as the probability that the binary classifier can distinguish between a randomly selected positive occurrence and a randomly selected negative occurrence.
S. pyogenes	*S. pyogenes* is a species of gram-positive bacteria. *S. pyogenes* plays a role in many human infections including NSTI. In the data used in this thesis, *S. pyogenes* was the most common cause of Type II NSTI infections.
S. dysgalactiae	*S. dysgalactiae* is a species of gram-positive bacteria capable of infecting humans and other animals. In the data used in this thesis, *S. dysgalactiae* was the cause of a small sub-set of severe NSTI patients.
SAPS II	In medicine, the simplified acute physiology score (SAPS II) is an ICU scoring system for determining the severity of disease and risk of mortality.
Septic shock	In medicine, septic shock is a condition that is defined by sepsis causing the occurrence of severely low blood pressure and abnormalities in the cellular mechanism. Sepsis is the damage or injury caused to an organ due to an infection.
Sigmoid curve	In mathematics, a sigmoid curve or function is a collection of functions that produce a "S"-shaped curve on the cartesion plane. The logistic function is an example of a sigmoid curve.
Signal Transduction	A common terminology used for the collection of biochemical cascades that use a series of proteins and molecular events to transmit signals in and out of cells.
SOFA score	In medicine, the sequential organ failure assessment score (SOFA score) is a scoring system in ICU that determines the rate of organ failure/loss of function in a patient.
SOP	Standard Operating Procedures in a clinical setting, are a set of written instructions for doing certain tasks.
ssPCC	ssPCC for Single Samples is a single-sample network inference method introduced by Liu et al.

Steady State	In dynamical systems, a system or a variable is said to be in a steady state when the value(s) of the system/variable do not change with respect to time. If the variable's change is modelled using an ODE, then the derivative of that variable with respect to time is 0. As an example a cytokine would be said to be in a steady state if the concentration of the cytokine doesn't change with the passing of time.
Thrombomodulin	Thrombomodulin is an integral membrane protein that is found on the surface of endothelial cells. Thrombomodulin has a negative feedback loop on Thrombin which leads to an anticoagulant effect. Thrombomodulin achieves this by forming the thrombin-thrombomodulin complex and activating protein C and protein S. In this thesis, Thrombomodulin was identified as a discriminatory plasma biomarker for necrosis identification.
Type I error	Type I error in statistics is the error in which there is an error in the rejection of the null hypothesis. This is also known as a false positive.
Type II error	Type II error in statistics is the error in which there is a failure to reject the null hypothesis. This is also known as a false negative.
Variable	Variable in statistics or feature in machine learning is a measurable piece of data that was used in the analysis.
Wilocoxon's test	Wilcoxon's rank sum test is another name for Mann-Whitney U test
Xgboost	Xgboost in machine learning is an algorithm similar to random forests. Xgboost uses an ensemble of decision trees and elements from gradient boosting algorithms to make its predictions. It sets itself apart from Random forests by pruning data splits that don't add to the improvement of the predictions and a computational ability to scale its process such that it is significantly faster than random forests.
Z-statistic test	In statistics, a Z-test is a group of tests in which the null hypothesis is defined such that the distribution of the test data can be approximated by a Gaussian (normal) distribution.

Chapter 2

Chapter



Sanjeevan Jahagirdar[1], Edoardo Saccenti[1]

Turn to page 377 for author affiliations

This chapter is adapted from:
Jahagirdar, S., & Saccenti, E. (2021). Evaluation of Single Sample Network Inference Methods for Metabolomics-Based Systems Medicine. ***Journal of Proteome Research*****,** ***20*****(1), 932–949. https://doi.org/10.1021/acs.jproteome.0c00696**

Personalising Metabolomics? A Closer Look at Single Sample Network Inference

Abstract

Networks and network analyses are fundamental tools of systems biology. Networks are built by inferring the pair-wise relationships among biological entities from a large number of samples such that subject specific information is lost. The possibility of constructing these sample (individual)-specific networks from single molecular profiles might offer new insights in systems and personalised medicine and as a consequence is attracting more and more research interest. In this study, we evaluated and compared LIONESS (Linear Interpolation to Obtain Network Estimates for Single Samples) and ssPCC (single sample network based on Pearson correlation) in the metabolomics context of metabolite-metabolite association networks. We illustrated and explored the characteristics of these two methods on *i*) simulated data, *ii*) data generated from a dynamic metabolic model to simulate real-life observed metabolite concentration profiles, *iii*) 22 metabolomic data sets and *iv*) we applied single sample network inference to a study case pertaining to the investigation of necrotizing soft tissue infections to show how these methods can be applied in metabolomics. We also proposed some adaptation of the methods that can be used for data exploration. Overall, despite some limitations, we found single sample networks to be a promising tool for the analysis of metabolomics data.

2.1 Introduction

Humans exhibit great phenotypic diversity in both healthy and pathophysiological conditions as a result of molecular regulatory and metabolic systems underlying the functioning of living organisms. It is now widely recognized that phenotypic diversity cannot be understood and characterized by analysing single molecular markers such as genes, metabolites or proteins alone: what is relevant is the complex web of interactions underlying the molecular mechanisms maintaining the functioning of the organism **(Futreal et al. 2004; Vidal et al. 2011; Zelezniak et al. 2014)**.

These molecular interactions are well captured and modeled using the formalism of network inference and analysis **(Barabasi et al. 2004; B. Zhang et al. 2014; Jinawath et al. 2016)**, where molecular entities such as genes, protein and metabolites are represented as nodes and their mutual relationships as edges, which can be different in nature, representing physical interactions as in protein-protein interaction networks, regulation as in gene regulatory networks, or similar concentration patterns as in metabolite-metabolite association networks **(Rosato, Tenori, Cascante, Carulla, et al. 2018)**.

It has been shown that network-based biomarkers, *e.g.* sub-network markers **(X. Liu, Z.-P. Liu, et al. 2012)**, network biomarkers **(R. Liu et al. 2014)** and edge biomarkers **(W. Zhang, T. Zeng, et al. 2015)** are superior to the traditional single-molecule biomarkers for accurately characterising disease states due to their additional information on interactions and networks.

In the quest for a personalized medicine **(Hamburg et al. 2010)** it is of paramount importance to elucidate the molecular mechanisms which underlay the subject-specific response to pathophysiological stimuli, resulting from the dysfunction of individual-specific networks/systems rather than just the malfunction of a singular biological entity. In this light, networks and network analyses have the potential of being pivotal in personalized medicine if there exists the possibility of their extension from a population level to an individual-specific level.

However, since several samples are required to define the associations (like in the form of correlations **(Suarez-Diez et al. 2015)**) among molecular elements like metabolites or genes, there exists no straightforward approach to infer an individual-specific network by profiling metabolite concentrations or gene expression from a single sample. If such an approach was demonstrated, it would be a very desirable situation due to the fact that it is rarely possible to obtain multiple samples from the same subjects, given the necessity of designing complex and expensive longitudinal studies. On the contrary, a single bio-fluid sample (such as blood and urine) is usually easy to obtain even in common clinical practice.

There is growing research interest in the possibility of the construction of such individual-specific networks by expression profiling of a single sample and several methods have been proposed **(X. Liu, Y. Wang, et al. 2016; Kuijjer, Tung, et al. 2019; Han et al. 2020; K. L. Buschur et al. 2020)**.

Here we present a comparative review of two methods for single sample network inference with the aim of evaluating their possible application to metabolomics data to obtain metabolite-metabolite single sample association networks. We focused on LIONESS proposed by Kuijjer et al. **(Kuijjer, Tung, et al. 2019)** and ssPCC by Liu et

al. **(X. Liu, Y. Wang, et al. 2016)**. We chose these two approach among others since they adopt similar albeit different philosophies and thus are directly comparable, are both (or may be) based on correlations and are easy to implement.

We analysed these two methods for their ability to produce single sample or sample-specific networks from metabolite concentrations and we explored and compared their characteristics to data generated from

- Numerical simulations
- A dynamic metabolic model
- 22 publicly available metabolomic data sets.

We then applied the two methods on a study case pertaining to the metabolomics investigation of nectrotising soft tissue infections (NSTI) **(Afzal et al. 2019)** in order to showcase the deployment of single sample network inference on real-life metabolomics applications. Additionally, we suggest some potentially new use of single sample networks for sample exploration and classification.

2.2 Methods

2.2.1 Basics of networks

A network is a graphical representation of relationships among objects. A network consists of nodes representing biological features (genes, proteins, metabolites) connected by links or edges which represent pair-wise relationships between the biological features.

This representation shifts the focus towards the relationships among biological entities rather than on their levels; in this light, network and network analysis are fundamental tools from the systems biology toolbox to investigate and understand metabolomics data **(Suarez-Diez et al. 2015)**. When the nodes are metabolites, the networks can be termed metabolite-metabolite association networks.

2.2.2 Methods for single sample networks inference

LIONESS: Linear Interpolation to Obtain Network Estimates for Single Samples

LIONESS is an approach developed by Kuijjer et al. in the context of gene regulatory networks **(Kuijjer, Tung, et al. 2019; Kuijjer, Hsieh, et al. 2019)**.

This approach starts by considering a $n \times m$ data matrix $\mathbf{X}_{(\alpha)}$ and the corresponding $m \times m$ network $\mathbf{E}^{(\alpha)}$ (*i.e* the so-called aggregate network) with edges e_{ij} between nodes i and j and the network $\mathbf{E}^{(\alpha-q)}$ constructed from the $(n-1) \times m$ data matrix $\mathbf{X}_{(\alpha-q)}$, that is a matrix with all but the q^{th} sample, which we refer to as the q-sample for sake of simplicity. A graphical illustration of the LIONESS procedure is given in Figure 2.1.

LIONESS assumes that the aggregate network $\mathbf{E}^{(\alpha)}$ built from n samples is the mean of networks constructed from every single sample from the data set $\mathbf{X}^{(\alpha)}$ containing n samples. This assumption is then extrapolated to define the edge $e_{ij}^{(\alpha)}$ in $\mathbf{E}^{(\alpha)}$

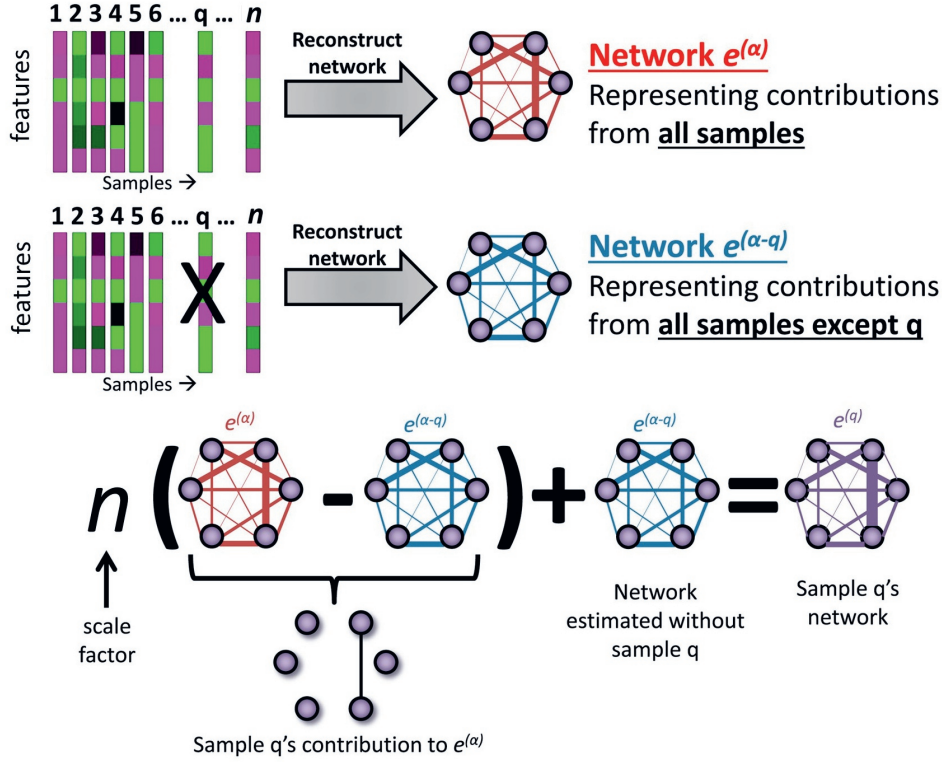


Figure 2.1: Visual illustration of the Lioness estimation of a single sample based on two aggregate network models, one reconstructed using all biological samples in a given data set and the other using all except the sample of interest (q, q-sample). Figure reproduced and adapted from the original publication **(Kuijjer, Tung, et al. 2019)**, published under Creative Commons license CC BY-NC-ND 4.0.

as the linear combination of the weights of that edge across a set of n networks:

$$e_{ij}^{(\alpha)} = \sum_{s=1}^{N} w_s^{(\alpha)} e_{ij}^{(s)}, \tag{2.1}$$

where $w_{(s)}$ represents the relative contribution of each single sample network to the aggregate network and

$$\sum_{s=1}^{N} w_s^{(\alpha)} = 1. \tag{2.2}$$

Similarly, for the network $\mathbf{E}^{(\alpha-q)}$ constructed from all but the q-sample, the edge between $e_{ij}^{\alpha-q}$ is defined as:

$$e_{ij}^{(\alpha-q)} = \sum_{s\neq q}^{N} w_s^{(\alpha-q)} e_{ij}^{(s)}, \tag{2.3}$$

where

$$\sum_{s \neq q}^{N} w_s^{(\alpha - q)} = 1.$$

From Equations (2.1) and (2.2), the authors defined

$$w_q^{(\alpha)} = 1 - w_s^{(\alpha)} / w_s^{(\alpha - q)}, \tag{2.4}$$

as long as the assumption holds that every sample makes equal proportional contribution to the aggregate networks $\mathbf{E}^{(\alpha)}$, which makes $w_q^{(\alpha)}$ constant.

Combining Equations (2.3) and (2.4) and solving for the edge $e_{ij}^{(q)}$ for the q-sample, gives the general LIONESS equation:

$$e_{ij}^{q} = \frac{1}{w_q^{(\alpha)}} \left(e_{ij}^{(\alpha)} - e_{ij}^{(\alpha - q)} \right) + e_{ij}^{(\alpha - q)}, \tag{2.5}$$

which define the edge between node i and j of the single sample network for the q-sample. The term $\frac{1}{w_q^{(\alpha)}}$ gives the weight of each sample, and can be set to n if all samples are given the same weight, obtaining

$$e_{ij}^{q} = n \left(e_{ij}^{(\alpha)} - e_{ij}^{(\alpha - q)} \right) + e_{ij}^{(\alpha - q)}, \tag{2.6}$$

which will be used throughout this study. In matrix notation Equation (2.6) becomes

$$\mathbf{E}^{q} = n \left(\mathbf{E}^{(\alpha)} - \mathbf{E}^{(\alpha - q)} \right) + \mathbf{E}^{(\alpha - q)}. \tag{2.7}$$

The LIONESS equation does not depend on the particular method used to infer the network, which can be estimated with any approach or different association measures: the most common approach is to use correlation, but mutual information can also be used. However, the authors reported sub-optimal performance when using mutual information to measure associations and single sample edges. In a related study, we also reported sub-optimal performance of mutual information for metabolite association network estimation **(Jahagirdar and Saccenti 2020b)**. For this reason, we will focus on Pearson's correlation, also because this will allow direct comparison with the ssPCC method (see Section *ssPCC: Single sample network based on Pearson's correlation*). In this case the LIONESS Equation (2.6) becomes

$$r_{ij}^{q} = n \left(r_{ij}^{(\alpha)} - r_{ij}^{(\alpha - q)} \right) + r_{ij}^{(\alpha - q)}, \tag{2.8}$$

where r_{ij} is the Pearson correlation between variable (metabolite) i and j. A summary of the notation used is given in Table 2.1.

Choice of the aggregate network

The LIONESS algorithm outputs a single sample network for each sample in a given data set given an aggregate network. If the data set contains n_1 samples from group

Table 2.1: Summary of the notation used in the paper to define the Lioness and ssPCC edges.

		Original notation	Definition	Alternative notation	Correlation notation	Definition
ssPCC	Network built using all samples in the reference data set	PCC_n			$r_{ij}^{(n)}$	
	Aggregated network built using all samples in the reference plus the q–sample	PCC_{n+1}			$r_{ij}^{(n+q)}$	
	Single sample network for the q–sample	ΔPCC_n	$PCC_{n+1} - PCC_n$		$r_{ij}^{(q)}$	$r_{ij}^{(n+q)} - r_{ij}^{(q)}$
Lioness	Aggregated network built using all samples	$e_{ij}^{(\alpha)}$		$e_{ij}^{(n)}$	$r_{ij}^{(\alpha)}$	
	Network built using all samples but the q–sample	$e_{ij}^{(\alpha-q)}$		$e_{ij}^{(n-q)}$	$r_{ij}^{(\alpha-q)}$	
	Single sample network for the q–sample	$e_{ij}^{(q)}$	$n(e_{ij}^{(\alpha)} - e_{ij}^{(\alpha-q)}) + e_{ij}^{(\alpha-q)}$		$r_{ij}^{(q)}$	$n(r_{ij}^{(\alpha)} - r_{ij}^{(\alpha-q)}) + r_{ij}^{(\alpha-q)}$

1 and n_2 samples from group 2, there is the legitimate question of whether to build the aggregate network using all $n_1 + n_2$ samples or to build two different separate reference networks, one for group 1 and one for group 2. In the original paper **(Kuijjer, Tung, et al. 2019)** the authors investigate the use of non-homogeneous background (page 13 of the supplementary material **(Kuijjer, Tung, et al. 2019)**) and reported minimal differences.

However, in this study we explored both implementations that we dubbed, for convenience as LIONESS single (LIONESS-S) and LIONESS double (LIONESS-D):

1. LIONESS-S Consider all samples to build the aggregated network and build single sample networks referring to the pool of all samples, or

2. LIONESS-D Consider two different aggregate networks $\mathbf{E}_1^{(\alpha)}$ and $\mathbf{E}_2^{(\alpha)}$ from the two groups samples and use them to build two sets single sample networks, one for group 1 and one for group 2.

ssPCC: Single sample network based on Pearson's correlation

The single sample network based on Pearson's correlation (which we abbreviate as ssPCC) was proposed by Liu et al. **(X. Liu, Y. Wang, et al. 2016)** for building sample-specific networks in the context of gene regulatory networks for disease characterisation. As such, it relies on the availability of a $n \times m$ $\mathbf{X}_n$ set of reference or control samples to which to contrast a set of case (possibly disease, in general from a different condition) q-samples. The ssPCC aims to define the single sample network specific to the q-sample(s).

Using the same notation used in the original publication, the single specific network for the q-samples obtained using as reference the n samples in $\mathbf{X}_n$, is given by

$$\Delta PCC_n = PCC_{n+1} - PCC_n \,, \tag{2.9}$$

where PCC_n is the Pearson's correlation matrix **(Pearson 1895)** calculated from the reference set $\mathbf{X}_n$ and PCC_{n+1} is the correlation matrix calculated from the $(n+1)\times m$ set made of $\mathbf{X}_n$ + the q-sample. The PCC_n is referred to as the "Reference network" while

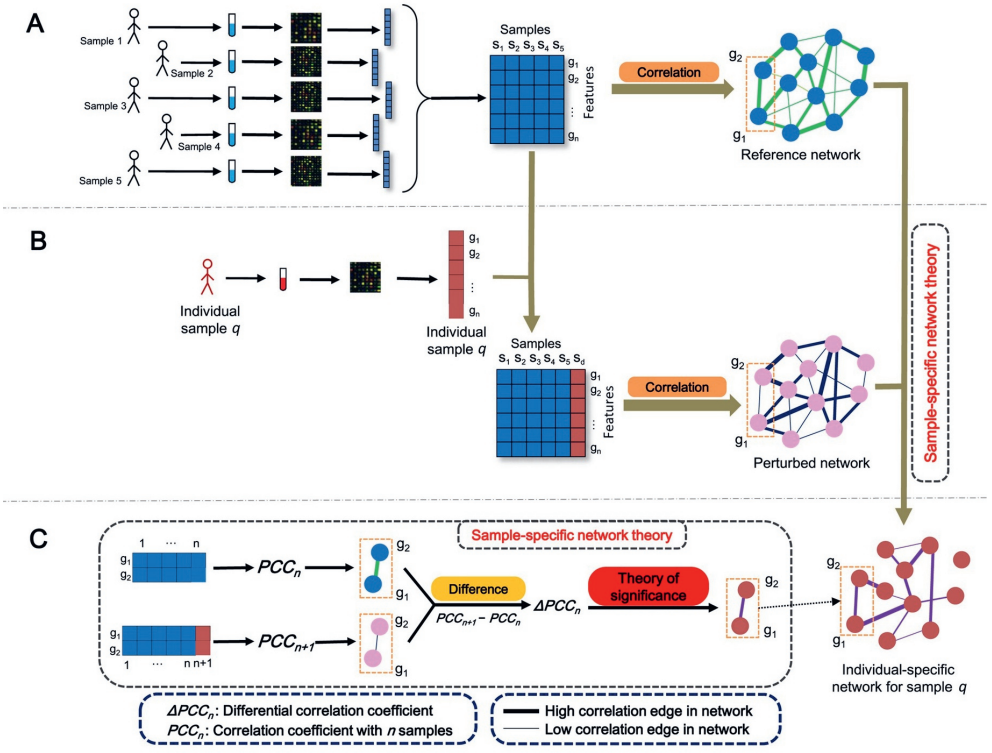


Figure 2.2: Visual illustration of the ssPCC estimation of single sample network. A) For a group of n reference samples, a reference network PCC_n can be constructed using Pearson's correlations. B) A new sample q is added and the perturbed network PCC_{n+1} with this additional sample is calculated. The difference between PCC_n and PCC_{n+1} is due to sample q. (C) The differential network ΔPCC_n is constructed taking the difference between perturbed and the reference network $PCC_n n+1 - PCC_n$. Figure reproduced and adapted from the original publication **(X. Liu, Y. Wang, et al. 2016)**, published under Creative Commons license CC BY-NC-ND.

PCC_{n+1} is referred to as the "Perturbed network". Thus the single sample network for the q-sample is considered to be the perturbation of the correlation of the m variables in $\mathbf{X}_n$ caused by the addition of the q-sample which comes from a different population. A graphical illustration of the ssPCC procedure is given in Figure 2.2.

Using an edge notation similar to the one used for LIONESS, the single sample network for the q-sample can be rewritten as

$$r_{ij}^{(q)} = r_{ij}^{(n+q)} - r_{ij}^{(n)}, \tag{2.10}$$

where the superscript $n+q$ indicates the addition of the sample q to the n samples of the reference matrix $\mathbf{X}_n$ and r_{ij} indicates the Pearson's sample correlation between variables i and j. A summary of the notation used is given in Table 2.1.

The authors proposed to assess the significance of an edge in the single sample network by means of permutation but found that the procedure could be conveniently

substituted with a Z-test which is much faster and gives equivalent results (see Section in the original paper **(X. Liu, Y. Wang, et al. 2016)**). They propose the following Z statistic test

$$Z = \frac{\Delta PCC_n}{(1 - PCC_n^2)/(n-1)}, \tag{2.11}$$

which, considering Equation (2.10), can be conveniently re-written using the edge notation in terms of the sample correlation between variable i and j

$$Z = \frac{r_{ij}^{(q)}}{1 - (r_{ij}^{(n)})^2/(n-1)}. \tag{2.12}$$

The Z-statistic is then confronted with the critical values of a standard normal distribution to assess significance.

ssPCC for a two group case

The ssPCC algorithm outputs single sample networks only for the case group and not for the reference group. This setting does not allow, *per se*, to build single sample networks for all samples (*i.e* case and reference samples) as in LIONESS. We attempted to bypass this limitation by building single sample networks also for the reference data set by contrasting each sample in the reference data set against the remaining samples, *i.e* considering each reference sample as a q-sample.

2.2.3 Data simulations

Numerical simulation scheme 1

We simulate $n \times 2$ reference data set $\mathbf{X}_n$ by sampling from a bivariate normal distribution

$$(x, y) \sim N(\mu_0, \Sigma_0), \tag{2.13}$$

with population $\boldsymbol{\mu}_0 = (0, 0)$ and

$$\Sigma_0 = \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}, \tag{2.14}$$

where ρ_0 is the population (expected) value of $r_{ij}^{(n)}$ (*i.e* of PCC_n) from ssPCC Equations (2.9) and (2.10).

We let ρ_0 to vary over the values $-0.9, -0.7, -0.5, -0.3, 0, 0.3, 0.5, 0.7$, and 0.9 to define 9 different reference covariance/correlation structures. The q-sample for which the single sample network is sought using ssPCC, *i.e* the sample to be added to $\mathbf{X}_n$ to calculate PCC_{n+1} is also drawn from a bivariate normal distribution with

$$(x, y)_q \sim N(\mu_q, \Sigma_q), \tag{2.15}$$

with population mean $\boldsymbol{\mu}_q = (0, 0)$ and correlation matrix

$$\Sigma_q = \begin{pmatrix} 1 & \rho_q \\ \rho_q & 1 \end{pmatrix}. \tag{2.16}$$

We let ρ_q to vary in the range $[0,+1]$ in increments of 0.1 to define 11 different covariance/correlation structures for variables x and y. We take the difference:

$$\delta = \rho_0 - \rho_q, \tag{2.17}$$

as a measure of the perturbation effect when a sample q is added to $\mathbf{X}_n$ to estimate $r_{ij}^{(n+q)}$ (*i.e* PCC_{n+1}): when $\rho_q = \rho_0$ the q-samples and the reference samples come from the same distribution, which implies that there is no perturbation, hence the expected value of $r_{ij}^{(n+q)}$ and $r_{ij}^{(n)}$ is the same. As ρ_q increases with respect to ρ_0, the perturbation increases, and in consequence ΔPCC_n also increases.

Numerical simulation scheme 2

This simulation is similar to the *Numerical simulation scheme 1*. The only difference is that

$$\boldsymbol{\mu}_q \neq \boldsymbol{\mu}_n, \tag{2.18}$$

i.e the q-samples come from a population with both different mean and correlation structure.

Numerical simulation scheme 3

We generated $m \times m$ (with $m = 20$) random correlation matrices Σ_m (with elements $\rho_{ij} \geq 0$ and $\rho_{ij} \neq \rho_{i'j'}$ for all possible variable pairs), satisfying the property

$$\frac{2}{m^2 - m} \sum_{i>j} |\rho_{ij}| = \rho, \tag{2.19}$$

This was achieved using the vine method **(Ghosh and Henderson 2003; Lewandowski et al. 2009)** by sampling from a Beta distribution $Beta(\alpha, \beta)$. The variance σ^2 of the Beta distribution was set to 0.1 and the mean μ was numerically optimised to have the sampled data obtain the required average correlation ρ equal to 0.1 to 0.9 in steps of 0.1, within a 5% precision. The mean μ and variance σ^2 are link the α and β parameters by the relationships

$$\begin{aligned} \mu &= \frac{\alpha}{\alpha + \beta} \\ \sigma^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \,. \end{aligned} \tag{2.20}$$

The optimised μ^{opt} values (0.113, 0.116, 0.123, 0.135, 0.163, 0.201, 0.262, and 0.382) were used to calculate the α and β shape parameters to be passed to the algorithm.

$$\begin{aligned} \alpha &= \frac{1}{\sigma^2} - \frac{1}{\mu^{opt}} \\ \beta &= \alpha\left(\frac{1}{\mu^{opt}} - 1\right). \end{aligned} \tag{2.21}$$

The correlation matrices Σ_m were used to simulate multivariate normally distributed data $N(\mu, \Sigma_m)$ to be used as reference and as q-samples. In particular, data with average correlation of 0.6 and 0.9 as reference data sets, data with average correlation in the range 0.1 to 0.9 were used as q-samples.

Multivariate data simulation

The covariance/correlation matrices defined in the *Numerical simulation scheme 1*, *Numerical simulation scheme 2*, and *Numerical simulation scheme 3* were used to generate normally distributed multivariate data using the Matlab mvnrnd function.

Data simulation using a dynamic metabolic model

To generate metabolites with plausible concentration patterns as observed in metabolomics data, we used a dynamic metabolic model, as described in the chapter *Correlation or Mutual Information? That is the Question!* **(Jahagirdar and Saccenti 2020b)**. For the sake of completeness, we report here the full simulation strategy.

The model describes the activation of NF-$\kappa\beta$ complex (nuclear factor kappa-light-chain-enhancer of activated B cells) and the corresponding response of the intracellular signaling pathway when exposed to lipo-polysaccharide that activates an inflammatory response. It consists of 59 ordinary differential equations recounting the reactions involving 35 metabolites. The model was obtained from the BioModels database **(Malik-Sheriff et al. 2020)** (www.ebi.ac.uk/biomodels/) with accession number BIOMD0000000489. Full details on the model building and accessory files can be found in the original publication. **(Sharp et al. 2013)**

The model dynamics were constructed and simulated using ordinary differential equations representing 3 types of reactions, namely,

- Reversible reactions using mass action law
- Irreversible reactions using mass action law
- Enzymatic reactions using Michaelis-Menten kinetics.

We give three reactions used in the model to showcase examples of ordinary differential equations corresponding to each type of reaction and the kinetics involved.

$$\frac{\mathrm{d}[\mathrm{IRF3[P]}]}{\mathrm{dt}} = kf_i \times [\mathrm{IRF3[P]}] - kr_i \times [\mathrm{IRF3[P](nuc)}], \tag{2.22}$$

$$\frac{\mathrm{d}[\mathrm{IRF3[P]}]}{\mathrm{dt}} = kf_i \times [\mathrm{IKK[P]}], \tag{2.23}$$

$$\frac{\mathrm{d}[\mathrm{IKK[P]}]}{\mathrm{dt}} = \frac{k_i \times [\mathrm{TAK1{:}TAB1{:}TAB2{:}TRAF6}] \times [\mathrm{IKK}]}{Km_i + [\mathrm{IKK}]}, \tag{2.24}$$

where IRF3 (Interferon Regulatory Factor 3), IKK(IkB kinase), TAK1:TAB1:TAB2:TRAF6 (complex of Mitogen-activated protein kinase kinase kinases) are three of the metabolites/compounds utilised in the $NF-\kappa\beta$ activation and corresponding signalling pathway, [P] represents the addition of Phosphoryl group via the process of phosphorylation, [metabolite] represents the concentration of the metabolites and i is the reaction number. For detailed description and information on all the metabolites and reactions involved in the $NF-\kappa\beta$ model, we refer the reader to the original publication and its Supplementary material. **(Sharp et al. 2013)**

Subject-specific concentration profiles were obtained by varying the kinetic constants Km_i, k_i, kf_i and kr_i for all of the 59 reactions and the initial concentrations c_m

for the 4 metabolites with non-null initial concentrations in order to generate subject-specific profiles. All these constants, Km_i, k_i, kf_i, kr_i and c_m were varied between bounds (a, b) ±10% of the original values (see Equation 2.25 and following) presented in the original publication. This was achieved by sampling from a uniform distribution $U(a, b)$ to obtain values for each subject. For the j-th individual, the values kf_i, kr_i,k_i, km_i, c_m for the i-th reaction were defined as

$$\begin{aligned} kf_i^j &\approx U(0.9 \times kf_i, 1.1 \times kf_i), \\ kr_i^j &\approx U(0.9 \times kr_i, 1.1 \times kr_i), \\ k_i^j &\approx U(0.9 \times k_i, 1.1 \times k_i), \\ Km_i^j &\approx U(0.9 \times Km_i, 1.1 \times Km_i), \\ c_m^j &\approx U(0.9 \times c_m, 1.1 \times c_m). \end{aligned} \tag{2.25}$$

The rationale of this approach is that models with different parameters will produce different metabolite profiles, like those observed when sampling different subjects in real life metabolomics experiments.

Using this approach we generated 500 individual profiles from which we build data sets of different size by random sampling.

To mimic different conditions we introduced perturbations to the model by manipulating the kinetic constants in the following manner:

$$\begin{aligned} \widetilde{kf}_i^j &= \epsilon \times kf_i^j, \\ \widetilde{kr}_i^j &= \epsilon \times kr_i^j, \\ \widetilde{k}_i^j &= \epsilon \times k_i^j, \\ \widetilde{Km}_i^j &= \epsilon \times Km_i^j, \\ \widetilde{c}_m^j &= \epsilon \times c_m^j. \end{aligned} \tag{2.26}$$

Here ϵ is used as a scaling parameter, the same for all reactions. The value of ϵ was varied over the values $\frac{1}{10}$, $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{1}{1.5}$, $1, 1.5, 2, 3, 5$, and 10 which were used to create subject-specific profiles in a similar manner as described above.

Using the same ϵ for all reactions allows us to investigate the performance of SSN method as function of the perturbation. From the pool of 500 samples we randomly sampled (with replacement) subsets of different size ($n = 10, 25, 50, 100, 250$ and 500).

2.2.4 Power of the ssPCC test

We investigated the actual power of the test (*i.e* the probability of rejecting the null hypothesis H_0 when actually false) by means of the *Numerical simulation scheme 1* and *Numerical simulation scheme 2*.

For each combination of Σ_n and Σ_q we generated a $n \times 2$ reference data set $\mathbf{X}_n$ from which we calculated PCC_n and k q-samples to obtain k values of ΔPCC_n which were tested for significance at the 0.05 level and recorded how many times H_0 was

correctly rejected to calculate the actual power of the test. In our Simulation scheme 1 the Null hypothesis H_0 is always false, except when $\Sigma_q = \Sigma_n$. The overall procedure was repeated 1000 times and the results were averaged over the repetitions. The actual power was calculated for $n = 25, 250$ and 25000.

The relative frequency of the rejection of H_0 when $\Sigma_q = \Sigma_n$ (*i.e.* H_0 true) is the actual α level of test, *i.e* the actual false positive rate (Type I error).

2.2.5 Principal component analysis

We explored differences among single sample networks with Principal component analysis (PCA). Each single sample matrix was vectorized and principal component analysis was applied on the edges to investigate the patterns of similarity/difference among the networks. Networks were vectorised by taking only the upper diagonal part of the network (given the symmetry), so that only $\frac{1}{2}m(m-1)$ edges are considered instead of m^2. Every $m \times m$ single-sample network was then collapsed to a $1 \times \frac{1}{2}m(m-1)$ vector, and the different networks were then collected in a matrix form suitable for PCA.

2.2.6 Random forest prediction models

Random Forest **(Liaw et al. 2002)** was used to building classification models to explore whether the use of single sample network edges in a prediction context, *i.e* to explore whether the use of the edge weights between pair of metabolites possess higher predictive power than the original metabolite concentrations. We focused on two-group scenarios, which are also the most commonly encountered in metabolomics applications, applying this approach to several public metabolomics data sets.

We built single sample networks using both LIONESS implementations (Single and Double, see Section *Choice of the aggregate network* and ssPCC (see Section *ssPCC: Single sample network based on Pearson's correlation*).

The single sample networks were processed for Random Forest as described in the case of PCA (Section *Data generation using a dynamic metabolic model*).

We used the standard Breiman's Random Forest implementation which uses the Gini impurity as loss **(Breiman 2001)**. We set the number of trees to 1000 and used the default value of $\sqrt{p}$ (where p is the number of variables) for the "mtry" parameter. We used a 2/3 + 1/3 data split (training + validation) to obtain an unbiased estimation of the classification. We took into account data unbalance using the "strata" option. Each model fitting was repeated 100 times to take into account the variability due to the re-sampling step used by the RF algorithm to randomly select the same number of subjects from each group and so to build the model on balanced data. The resampling was nested within the cross-validation step used to assess the quality of the prediction models. All results are given as the arithmetic mean over the 100 iterations.

2.2.7 Pathway enrichment analysis

Pathway enrichment analysis was performed using the built-in function available in the MetaboAnalyst 4.0 **(Chong et al. 2018)** online server (www.metaboanalyst.ca) using the hypergeometric test. The Benjamini-Hochberg method was used for false

discovery rate (fdr) correction **(Benjamini et al. 1995)**. We considered pathways with fdr < 0.01 as significantly enriched.

2.2.8 Experimental data

Metabolomic study case

As a study case, we considered a data set from a metabolomic investigation of necrotising soft tissue infections (NSTI). The data set consists of plasma metabolite profiles acquired via GC-MS on 34 NSTI patients enrolled in the INFECT project (href-Clinicaltrials.govClinicaltrials.gov, NCT01790698). In addition, 24 patients with no known infections were included as controls.

The patients had NSTI of different microbial aetiology and were classified into polymicrobial and monomicrobial NSTI.

This data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, (www.metabolomicsworkbench.org) with Project ID ST00127 where it can be accessed via the Project DOI (DOI: 10.21228/M85H5H). We refer to the original publication **(Afzal et al. 2019)** for more details on the study design, sample collection and processing, GC-MS experiments and clinical information.

Compendium of publicly available metabolomics data

To further explore the characteristics of the two methods for inference of single sample networks, we used a compendium of 22 data sets that we compiled for a previous study **(Jahagirdar and Saccenti 2020b)**.

Ten of these data sets were derived from the compendium assembled by Mendez et al. **(Mendez et al. 2019)** including studies representatives of three of the most frequent metabolomic experimental platforms (nuclear magnetic resonance, NMR; gas chromatography mass spectrometry, GC-MS; liquid chromatography mass spectrometry, LC-MS) concerning the metabolomic profiling of varying biofluids (serum, plasma, urine, faeces). All studies follow a two-group design (either as primary design or secondary result of the publication, or as a subset of a multi-class study) and have varying sample sizes and number of metabolites (variables) present. We have made use of the processed and cleaned data available in Mendez et al. **(Mendez et al. 2019)** to which we refer the reader for more detailed information on the processing and cleaning of the data sets. Metabolites having missing values were either deleted (*e.g.* dataset MTBLS136) or imputed using the Random Forest procedure implemented in the R package missForest **(Stekhoven et al. 2011)** (*e.g.* data set ST001047). We also included other data sets derived from tissue (fat), plant and fruit extracts along with microbiome data (16S sequencing) and other chemical based assays on various fluids like coffee, wine and oil and finally two transcriptomic data sets. Relevant references and attributes for all data sets are shown in Table 2.2.

2.2.9 Software

Calculations were performed using R **(R Core Team 2013)**, Matlab **(MATLAB 2018)** and Python **(Python Core Team 2015)**. Our R implementations for LIONESS and

Table 2.2: Random Forest classifications of 22 case-control metabolomics data sets using metabolite concentrations and single sample network edges as described in Section *Experimental data*. Abbreviations: CD Chron's disease. UC Ulcerative colitis. Assay stands for chemical assay.

	Study ID	Ref	Platform	Type	Obs	Var	Design	Classification Accuracy			
								Conc	ssPCC	Lioness-S	Lioness-D
1	KODAMA	**(Bernini et al. 2009)**	NMR	Urine	80(40/40)	490	Subject (A/B)	96.9	100.0	87.5	100.0
2	MTBLS123	**(Lusczek et al. 2013)**	NMR	Urine	151 (79/72)	63	Shock (pre/post)	99.1	97.4	56.3	81.5
3	MTBLS136	**(V. L. Stevens et al. 2018)**	LC-MS	Serum	668 (337/331)	371	Postmenopausal hormone (estrogen/estrogen+proges	99.1	99.6	59.4	89.8
4	MTBLS161	**(C. W. Armstrong et al. 2015)**	NMR	Serum	59 (34/25)	30	Chronic fatigue syndrome (case/control)	96.0	96.6	67.8	86.4
5	MTBLS404	**(Thévenot et al. 2015)**	LC-MS	Urine	184 (101/83)	120	Sex (M/F)	98.3	100.0	67.4	100.0
6	MTBLS547	**(X. Zheng et al. 2017)**	LC-MS	Caecal	97 (46/51)	35	High fat diet (case/control)	99.9	97.9	84.5	94.8
7	MTBLS90	**(Ganna et al. 2014)**	LC-MS	Plasma	968 (485/483)	189	Sex (M/F)	99.4	90.8	63.6	91.3
8	MTBLS92	**(Hilvo et al. 2014)**	LC-MS	Plasma	253 (142/111)	138	Breast cancer chemotherap (before/after)	98.3	92.5	62.7	89.3
9	pgmm	**(Forina, Armanino, Lanteri, et al. 1983)**	Assay	Oil	50 (25/25)	7	Region (A/B)	100.0	97.3	92.0	98.7
10	pgmm	**(Streuli 1973)**	Assay	Coffee	43 (36/7)	12	Variety (Arabica/Robusta)	100.0	97.7	95.3	83.7
11	pgmm	**(Forina, Armanino, Castino, et al. 1986)**	Assay	Wine	130 (59/71)	27	Type (Barolo/Grignolino)	100.0	100.0	81.5	94.6
12	ST000061		GC-MS	Tissue	118 (59/59)	157	subcutaeus/visceral fat	94.7	99.1	78.6	87.2
13	ST000369	**(Fahrmann et al. 2015)**	GC-MS	Serum	80 (49/31)	181	Adenocarcinoma (case/-control)	89.9	100.0	55.0	80.0
14	ST000496	**(Sakanaka et al. 2017)**	GC-MS	Saliva	100 (50/50)	69	Debridement (pre/post)	99.3	96.0	63.0	91.0
15	ST001000	**(Franzosa, Sirota-Madi, et al. 2019)**	LC-MS	Stool	121 (68/53)	124	Inflammatory bowel diseas (CD/UC)	91.7	98.3	66.9	96.7
16	ST001047	**(A. W. Chan et al. 2016)**	NMR	Urine	83 (43/40)	149	Gastric cancer (gastric cancer/healthy)	93.4	100.0	61.4	88.0
17	ST001243	**(Powers et al. 2019)**	GC-MS	Plasma	98 (48/50)	69	Trisomy 21 (yes/no)	99.0	100.0	79.2	91.7
18		**(Eisner et al. 2011)**	NMR	Urine	50 (25/25)	200	cachexia (case/control)	92.2	93.5	70.1	94.8
19		**(Eisner et al. 2011)**	NMR	Urine	77 (47/30)	63		94.4	94.0	96.0	98.0
20		**(Eisner et al. 2011)**	NMR	Urine	60 (30/30)	63		99.4	96.2	67.4	86.6
21		**(Rist et al. 2017)**	GC-MS	Urine	301 (129/172)	324	Sex (M/F)	98.2	99.0	72.4	90.0
22		**(Caldana et al. 2011)**	GC-MS	Plant	70 (35/35)	67	Light/Dark	91.5	94.3	62.9	84.3

ssPCC are available at www.systemsbiology.nl under the software tab. Original R package for LIONESS by Kuijer et al. **(Kuijjer, Hsieh, et al. 2019)** can be also obtained at https://github.com/kuijjerlab/lionessR and
https://bioconductor.org/packages/release/bioc/html/lionessR.html.

2.3 Results and Discussion

We begin by noticing that LIONESS and ssPCC are not context or data dependent or depending on how the networks are inferred. Both methods have been originally applied to gene regulatory networks but the statistical framework is totally general: they are both based on "manipulation" of correlations but the way the correlations are calculated and manipulated is totally independent from their origin. There is nothing in how the methods are formulated that is specifically depending on or descending from the correlations originating from gene regulation patterns. The two frameworks are fully generalisable to different biological contexts and applications. Here we explore their applicability to metabolite-metabolite correlation networks: different approaches can be used to calculate the reference networks depending on the applications but the way the single-sample networks are obtained does not depend on the application.

2.3.1 Power of the ssPCC test

We investigated the actual power of the ssPCC test (*i.e* the probability of rejecting the null hypothesis when actually false) using the simulation scheme 1 described in the Material and Methods section.

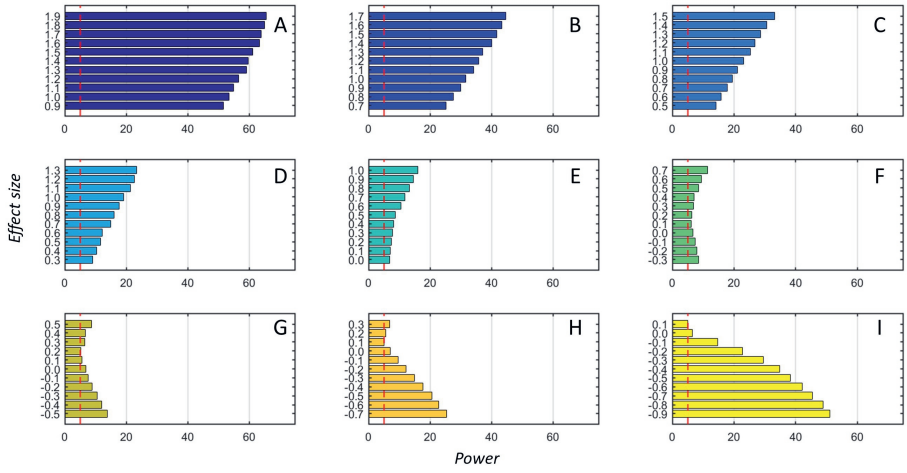


Figure 2.3: Actual power of the ssPCC proposed testing procedure based on a Z-test (see Section *ssPCC: Single sample network based on Pearson's correlation* and Equation 2.12) as function of the effect size defined in Section *Power of the ssPCC test* using the simulation scheme 1. The nine panels A to I correspond to the different correlation ρ_0 of the ssPCC reference network (PCC_n, see Equation 2.14): the values are $-0.9, -0.7, -0.5, -0.3, 0, 0.3, 0.5, 0.7$, and 0.9 respectively. For each value of ρ_0, we let ρ_q (see Equation 2.16) to vary in the range [0:1] with increments of 0.1. The effect size is defined as $\rho_q - \rho_0$, so that for each ρ_0 there are 11 effect size values, indicated on the $y-axis$ of the bar plots. Results are shown for sample size $n = 1$. The actual power is calculated over 1000 tests. The vertical dashed red lines indicate the 5% level.

As shown in Figure 2.3 we observed limited power, when the reference samples and the q-samples are drawn from a multivariate normal distribution with different population covariance-correlation matrices but with the same population means (Simulation 1). The results shown are for $n = 250$.

The power of the test is heavily dependent on the correlation among x and y in the reference set, *i.e* on PCC_n: the maximum power achieved is slightly higher than 60% when $r_{ij}^{(n)} = 0.9$ (PCC_n) (panel A) and reaches its minimum when $r_{ij}^{(n)}$ is below 0.5 (panels E, F and G).

Unfortunately this is the most common case in metabolomics studies: metabolomics correlations are mostly positive and smaller than 0.6, as also shown in Table 2.3, where the distribution characteristics of correlation values (*i.e* $r_{ij}^{(n)}, PCC_n$ are given for 10 case-control metabolomics data sets. Thus the most common situation observed in real life metabolomics studies is the situation in which the test has less power to detect differences between the q-sample and the reference set.

The power of the test depends of course on n, and should tend to 1 as $n \to \infty$. However we did not observe any strong dependence on n: for $n = 25$ the maximum power was $\approx 64\%$ which increased to $\approx 66\%$ for $n = 25000$ when $PCC_n = 0.9$, which indicates that the size of the reference data set has little influence on the actual power

Table 2.3: Association between metabolite differential expression and single sample edges found to be significant with the proposed ssPCC test. The % of edges associated with differential expressed metabolites is shown together with the average ssPCC edges and with the average correlation of the reference network calculated on 10 case-control metabolomics data sets. More information on the data can be found in Table 2.2

		% SSN Edges associated with DE metabolites			Average ssPCC SSN edge			Correlation in reference Data		
	Data set	Mean	95% CI		Mean	95% CI		Mean	95% CI	
1	MTBLS136	56.1	39.6	77.8	0.01	0.01	0.01	0.03	-0.17	0.43
2	MTBLS161	82.3	42.6	98.3	0.02	0.00	0.12	0.35	-0.38	0.79
3	MTBLS404	67.9	0.0	97.3	0.01	0.00	0.05	0.30	-0.05	0.73
4	MTBLS547	76.3	0.0	100.0	0.01	0.00	0.07	0.29	-0.22	0.84
5	MTBLS90	84.7	69.3	97.8	0.01	0.00	0.01	0.06	-0.18	0.62
6	MTBLS92	85.0	65.3	98.3	0.01	0.00	0.03	0.21	-0.09	0.80
7	ST000369	39.5	14.6	68.1	0.11	0.08	0.17	0.03	-0.31	0.50
8	ST000369	54.1	32.0	77.8	0.02	0.01	0.08	0.03	-0.29	0.45
9	ST001000	75.3	0.0	98.2	0.02	0.00	0.09	0.14	-0.28	0.95
10	ST001047	75.3	0.0	98.2	0.02	0.00	0.09	0.14	-0.28	0.95

		% SSN Edges associated with DE metabolites			Average Lioness SSN edge			Correlation in aggregate Data		
	Data set	Mean	95% CI		Mean	95% CI		Mean	95% CI	
1	MTBLS136	55.10	49.1	62.4	0.55	0.33	0.95	0.03	-0.16	0.43
2	MTBLS161	80.02	42.7	88.2	0.63	0.34	1.22	0.30	-0.19	0.67
3	MTBLS404	79.34	62.3	87.5	0.60	0.32	1.20	0.24	-0.05	0.66
4	MTBLS547	81.20	63.9	88.6	0.62	0.33	1.22	0.31	-0.13	0.77
5	MTBLS90	82.46	76.5	87.8	0.63	0.34	1.16	0.06	-0.17	0.62
6	MTBLS92	82.92	76.5	88.8	0.63	0.34	1.23	0.19	-0.07	0.80
7	ST000369	31.98	23.7	42.6	0.59	0.32	1.88	0.04	-0.23	0.60
8	ST000369	51.28	35.7	61.5	0.54	0.32	0.95	0.03	-0.22	0.53
9	ST001000	81.10	65.2	87.8	0.60	0.33	1.10	0.15	-0.25	0.93
10	ST001047	81.10	65.2	87.8	0.60	0.33	1.10	0.15	-0.25	0.93

of the test.

The limited power of the test under simulation scheme 1 can be understood by considering Equation 4 in the original publication.

The authors derived an interesting relationship (in the case $n >> 1$) linking the ΔPCC_n, *i.e* $r_{ij}^{(q)}$ and the difference of the level of variable i and j measured on the q-sample with respect to the average level of the same variables in the reference data:

$$\Delta PCC \approx \frac{1}{n-1}\left(\Delta x \Delta y \; - \frac{PCC_n}{2}(\Delta x^2 + \Delta y^2)\right), \tag{2.27}$$

where

$$\Delta x = \frac{x - m_x^{(n)}}{\sigma_x^2}. \tag{2.28}$$

Several interesting observations can be derived from Equation (2.27):

1. $\Delta PCC_n \to 0$ if Δx and Δy are zero, that if the q sample is from a population with same average level ($\boldsymbol{\mu}_q = \boldsymbol{\mu}_n$) of the reference population as in Simulation 1.

2. PCC_{n+1} does not appear in the (re)definition of ΔPCC_n: only the difference in the levels of X and Y with respect to the correlation of X and Y in the reference define ΔPCC_n.

This explains the very limited power observed in Figure 2.3. If the reference samples and the q-samples are drawn from multivariate distribution with the same population means both Δx and Δy tend to 0 and ΔPCC_n tend to zero even if $PCC_n \neq 0$.

3. If the number of samples in the reference data set n is very large, adding the q sample has practically no influence: it is not the perturbation of the correlation (*i.e* $PCC_{n+1} - PCC_n$) that it is tested, but some function of the differences of X and Y with respect to the average values in reference data. This observation is supported by empirical evidence shown in Table 2.3 where SSN networks were built for 10 case-control metabolomics data sets: the vast majority of significant SSN edges are associated with metabolites whose concentrations are significantly different between the two conditions.

4. If we plug Equation (2.27) in Equation (2.11) we obtain the following expression for the Z-statistic
$$Z = \frac{\Delta x \Delta y \ - \frac{1}{2} PCC_n(\Delta x^2 + \Delta y^2)}{1 - PCC_n^2}, \tag{2.29}$$
which does not depend explicitly on n: since PCC_n is fixed *a priori*, this explains why increasing the dimensionality of the reference data set $\mathbf{X}_n$ has little influence on the power of the test.

5. ΔPCC_n can be different from zero also when PCC_n is zero, that is when the reference samples and the q-samples are from populations with the same covariance-correlation structure: this happens when $\Delta x, \Delta y \neq 0$. This explains the slightly inflated Type I error observed in Figure 2.4.

6. ΔPCC_n can be zero even if $\Delta x, \Delta y$ and PCC_n are all different from zero. This can happen, for instance, when
$$\begin{aligned} \Delta x &= \sin(\frac{1}{2}\arcsin(PCC_n)), \\ \Delta y &= \cos(\frac{1}{2}\arcsin(PCC_n)), \end{aligned} \tag{2.30}$$
or
$$\begin{aligned} \Delta x &= \sin(\frac{1}{2}\arcsin(r_{ij}^{(n)}), \\ \Delta y &= \cos(\frac{1}{2}\arcsin(r_{ij}^{(n)})). \end{aligned} \tag{2.31}$$

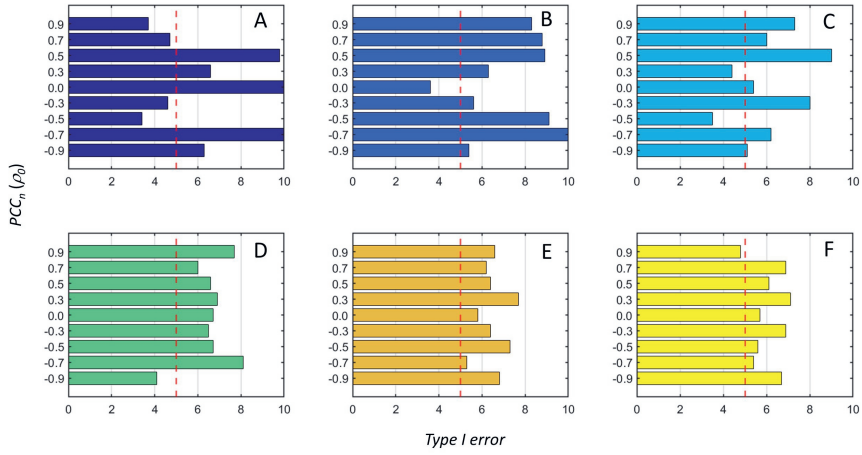


Figure 2.4: Actual Type I error (α) of the ssPCC proposed testing procedure based on a Z-test (see Section *ssPCC: Single sample network based on Pearson's correlation* and Equation 2.12) as function of the effect size defined in Section *Power of the ssPCC test*. The nominal 5% level is indicated by the vertical dashed red line. The six panels A to F corresponds to the different sample size used, with 10, 50, 100, 500, 1000, and 10000, respectively. The actual α is calculated over 1000 tests.

2.3.2 Relationship between LIONESS and ssPCC single sample networks

If correlation is used as a measure of variable association, the two methods are functionally related, and from the definition of the ssPCC and LIONESS edges it follows that the latter can be written as a function of the ssPCC edges. In particular, when the data set $\mathbf{X}^{(\alpha)}$ used to build the aggregated network in LIONESS is equivalent to the reference data set used in ssPCC, their is an almost perfect linear relationship between the edges of the q-sample network estimated using LIONESS and using ssPCC, as it can be seen in Figure 2.5 panel A. This relationship deteriorates when $\mathbf{X}^{(\alpha)}$ is not equal to X_n, that is when there is more than one sample belonging to a group different from the reference (Figure 2.5 panel B). Note the different scale of the edge weights: for ssPCC the edges are bounded between -2 and +2 being defined as the difference between two correlations; for LIONESS, if correlations are used, the edges are bounded between $1-2n$ and $2n-1$.

A relationship similar to Equation (2.27) can be derived also for LIONESS. It is enough to note that $r_{ij}^{(n+1)}$ is actually the correlation calculated using all samples (thus including the q-samples *i.e* $r_{ij}^{(\alpha)}$) in LIONESS and PCC_n is the correlation calculated

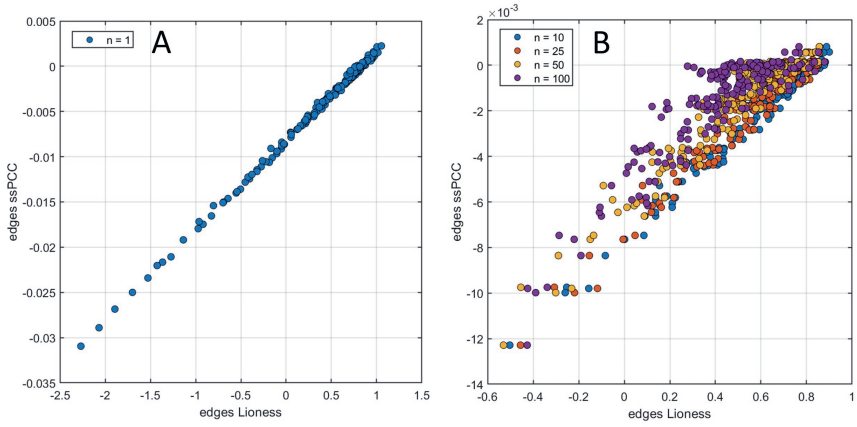


Figure 2.5: Comparison of the Lioness and ssPCC single sample edges. A) scatter plot of the edges for a q-sample obtained when the data set $\mathbf{X}^{(\alpha)}$ used to build the aggregated network in Lioness is equivalent to the reference data set used in ssPCC. B) Scatter plot of the edges for a q-sample obtained when $\mathbf{X}^{(\alpha)}$ is not equal to $\mathbf{X}_n$, *i.e* when there is more than one sample belonging to a group different from the reference. Note the different scale of the edge weights.

using all but the q-sample, *i.e* $r_{ij}^{(\alpha-q)}$. The LIONESS edge can be re-written as

$$r_{ij}^{(q)} \approx \frac{n}{n-2}\left(\Delta x \Delta y - \frac{r_{ij}^{(\alpha-q)}}{2}(\Delta X_i^2 + \Delta x_j^2)\right) + r_{ij}^{(\alpha-q)}. \tag{2.32}$$

This establishes that also the edges of a single sample network estimated with LIONESS are a function of both the perturbation of the correlation and of the difference between the levels of the q-sample and the mean of the remaining $n-1$ samples.

2.3.3 Comparison of ssPCC and LIONESS on simulated data

Simulated data with different levels of correlation

The first simulation entails the case of the reference data ($m = 20$ variables) with zero mean and average correlation equal to 0.6: this value was chosen because most part of the observed metabolomic correlations are smaller than 0.6 **(D. Camacho et al. 2005)**. The q-sample comes from data with average correlation between 0 and 0.9 and mean equal to 0 or 10. The results are shown in Figure 2.6, panels A to D. As it can be seen there is no obvious separation of the single sample networks as function of the effect size, although a slight separation appears when the mean of the reference and q-sample differ, especially between single sample networks from sample with extreme average correlation (*i.e* 0 and 0.9).

When the same exercise is repeated with reference data with average correlation equal to 0.9 (Figure 2.6, panel E to H), results are similar, with a slight separation

emerging only when the reference samples and the q-samples come from populations with different means.

Comparison on $NF-\kappa\beta$ dynamic model data

We compared the LIONESS and ssPCC on data simulated from a dynamic model for the $NF-\kappa\beta$ pathway where model parameters were manipulated to introduce increasing levels of perturbation with respect to the original unperturbed model: this was accomplished by multiplying/dividing the kinetic parameter (see Equations (2.27 with $\epsilon = 1, 1.5, 2, 3, 5,$ and 10). Results are shown in Figure 2.7; $n = 50$ samples were considered for each configuration.

In the case of the multiplicative perturbation of the model, networks tend to cluster according to the level of perturbation, with highly perturbed networks clustering away from those corresponding to low perturbation. This is particularly evident for ssPCC derived networks, with a very clear separation among the clusters (Figure 2.7 panel B); however, the separation among LIONESS based networks is much less evident (Panel A).

For data from the perturbed model obtained by dividing the kinetic constants (see Equation 2.27 with $\epsilon = \frac{1}{10}, \frac{1}{5}, \frac{1}{3}, \frac{1}{2},$ and $\frac{1}{1.5}$), the single sample networks obtained with LIONESS are not resolved and it is not possible to distinguish among the different groups corresponding to the different perturbation levels (Figure 2.7, panel C). On the contrary the networks obtained with ssPCC are very well resolved, and clear differences appear among the groups (Figure 2.7, panel D).

2.3.4 Use of single sample edges for group prediction and classification

We explored the potential of single sample network edges for classification purposes by replacing actual observed metabolite concentrations with the pairwise edges, the rationale being that, as follows from Equations (2.27) and (2.32), the single sample edges are a function of both difference in correlation and in level and thus, in some case, can bear more information than level alone.

To this scope, we compared the accuracy of Random Forest classification models on 25 publicly available data sets (see Section *Compendium of publicly available metabolomics data*). For each data set, we built four different Random Forest classification models using:

1. Original concentration/abundance profiles.
2. The edges of the single sample networks built using ssPCC.
3. The edges of the single sample networks built using LIONESS and all samples to build the aggregrate network (LIONESS-D).
4. The edges of the single sample networks built using LIONESS and only group specific samples to build the aggregrate network (LIONESS-S).

In total we have three ways to build single sample networks edges for a two-class problem to be used for classification purposes.

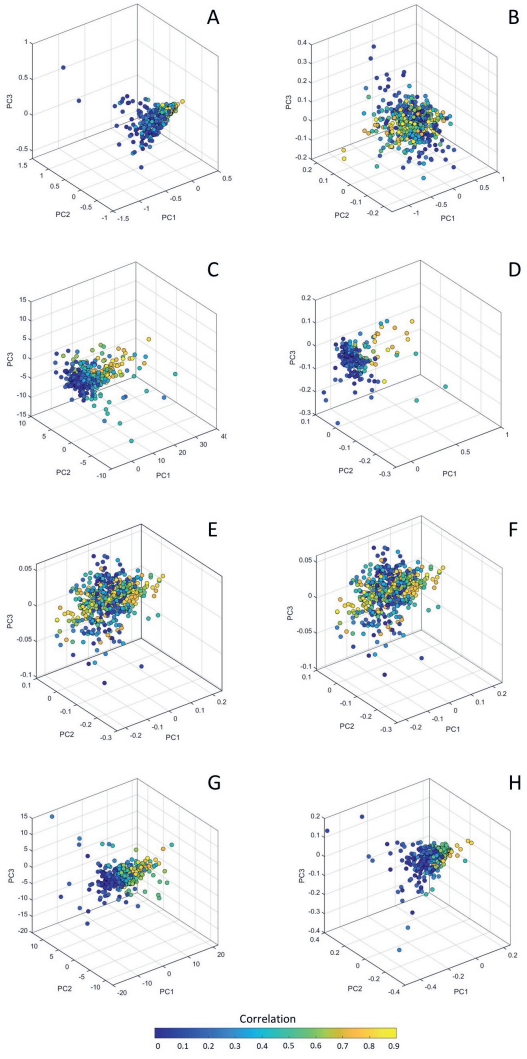


Figure 2.6: Scatter plot from Principal Component Analysis of the single sample networks obtained with Lioness-S (left column) and ssPCC (right column). Panels A to D refer to simulation with reference data (20 variables) with average correlation equal to 0.6 while the q-sample is from a population with correlation between 0 and 0.9 and mean equal to 0 (Panels A and B) or 10 (Panels C and D). Panels E to H refer to simulation with reference data with average correlation 0.9 while the q-sample is from a populations with correlation between 0 and 0.9 and mean equal to 0 (*i.e* data from simulation 1, Panels E and F) or 10 (*i.e* data from simulation 2, Panels G and H). Each point in the PCA plot is a vectorized version of the single sample networks, colour coded according the population correlation from which the q-samples are sampled.

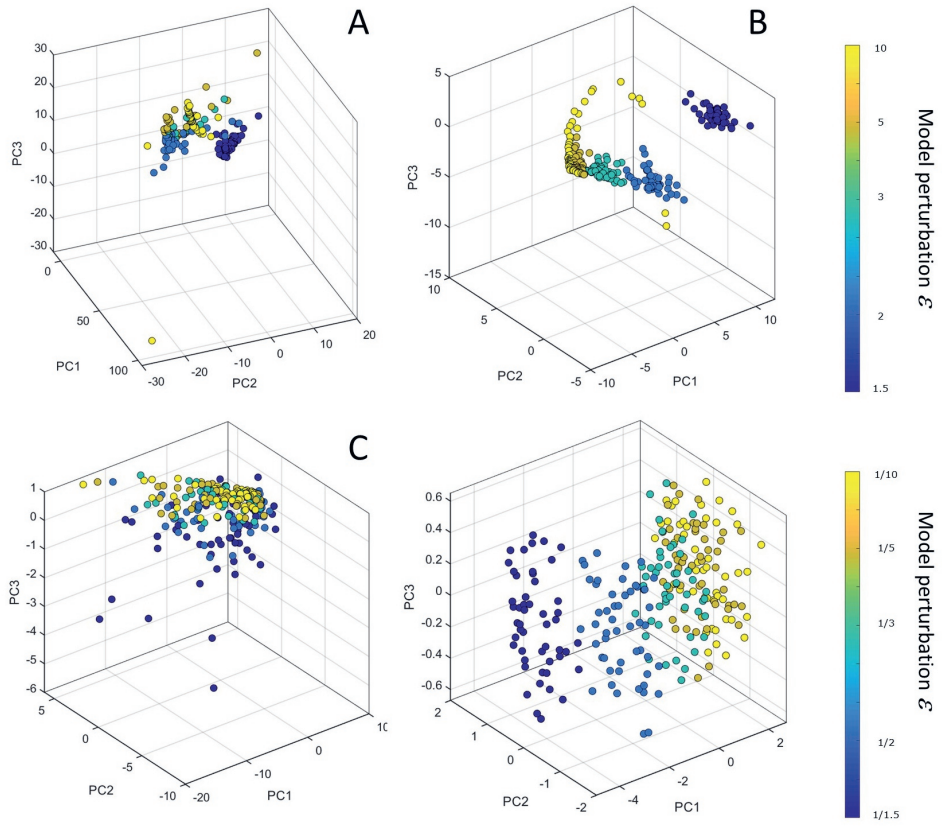


Figure 2.7: Scatter plot from Principal Component Analysis of the single sample networks obtained with Lioness-S (left column) and ssPCC (right column) on data simulated using a dynamic model of the NF-kB metabolic pathway. Panels A and B given results on data obtained by multiplying the kinetic parameters by a factor $\epsilon = 1, 1.5, 2, 3, 5$, and 10, see Equations 2.27. Panels C and D give results on data obtained by multiplying the kinetic parameters by a factor $\epsilon = \frac{1}{10}, \frac{1}{5}, \frac{1}{3}, \frac{1}{2}$, and $\frac{1}{1.5}$). Each point in the PCA plot is a vectorized version of the single sample networks, colour coded by the magnitude of the perturbation. For each ϵ configurations $n = 50$ samples where considered.

The results are given in Table 2.2. Under the assumption that single sample edges hold more information content than the concentration profiles, we expected the Random Forest models built on edges to have higher accuracy than those built on the original concentration values.

We observed that in general, classification accuracy is already very high when using the original values. However, in 13 cases out of 22, the use of ssPCC edges gives better (or equal) classification accuracy when ssPCC edges are used.

For some data sets the use of single-sample edges in place of concentrations resulted in better discrimination: for instance, for data set 13 the accuracy increases from 89.9% to 100% and better discrimination between cases and controls is also obtained for data sets 1, 3, 4, 5, 12, 15, 16, 17, 18, 20 and 22. In other cases a reduction in accuracy can be observed.

When using the LIONESS-D approach (akin to the strategy we devised for ssPCC) the accuracy was better only in 5 cases out of 22, while using the LIONESS-S approach the use of single sample networks was better only in one case: in all other cases the accuracy was remarkably lower.

When Random Forest was run on edges from LIONESS-S, discrimination accuracy was lower with the sole exception of data set 19. This can be explained by the process of separating case and control data before running the LIONESS algorithm. In this format, the matrix E^q is built as a function of $E^{(\alpha)}$ as shown in equation 2.7, however here all the samples in $E^{(\alpha)}$ come from the same classification group causing the variation between single samples to be much lower due to the lower difference in E^{α} and $E^{(\alpha-q)}$ as compared to LIONESS-D.

There is a large difference depending on how the aggregate network is built. In the original publication **(Kuijjer, Tung, et al. 2019)** the authors discussed the problem of how to build the aggregate network when in presence of samples from non-homogeneous populations (basically considering the two approaches used here). They found minimal difference (verbatim) in the accuracy of the reconstruction of the single sample networks but did not explore the edge reconstruction in a discriminant/classification setting.

We can also comment that calculating correlations from a non-homogeneous population is not statistically a sound strategy, since sample correlation must be calculated from samples drawn from the same population. It is simple to show with simulations that if half of the samples come from a population with correlation $\rho = 1$ and a half from a population with $\rho = 0$, the sample correlation will be around 0.5 and this will lower the single sample edge.

Classification models built from ssPCC and LIONESS-D edges often yielded similar albeit lower accuracy. This is not too unexpected as when LIONESS is expressed as a function of ssPCC as shown in equation 2.32 from which it descends that the LIONESS single sample edge also depends on Δx_i and Δx_j which are the deviation of the q-sample from the average of the reference data: in the LIONESS-D approach, the deviations Δx_i and Δx_j are calculated from each q-sample from data sets that are homogeneous to the q-sample and thus can be expected to be small, and this also lowers the values of the edges. This is not the case for all the ssPCC edges: in our modification, the deviations for q-samples in the control group are estimated from samples homogeneous to the q-samples but for the control group the q-samples are not homogeneous to the reference (which is made from control samples) and

this makes the edges larger. This can explain the markedly different behavior of the different single sample edges when used for discrimination between two groups.

We shall conclude by remarking that in a classification setting the use of single sample edges as derived from ssPCC and LIONESS-D can be used only in an exploratory or confirmatory setting but not to predict new, unknown, q-samples: the network must be constructed by contrasting the unknown sample either with the case or the control group without knowing to which group the q-sample actually belongs. Unknown q-samples can be predicted using the edges from LIONESS when the aggregates network is constructed using all the samples from both groups simultaneously.

2.3.5 Metabolomics case study: Necrotising soft tissue infections

In order to delve deeper into the characteristics of the two methods and to investigate whether (possibly) new biological information can be gained from the use of single sample networks, we analysed in detail metabolomics data concerning metabolite plasma profiles collected from patients suffering from necrotising soft tissue infections, fast spreading, aggressive bacterial infections associated with a high morbidity and mortality **(Anaya et al. 2005; D. L. Stevens and Bryant 2017)**. The study comprised 34 NSTI patients and 24 surgery patients with no known infection or morbidity acting as controls. This data has been previously analyzed using standard statistical (univariate) approaches and differential correlation analysis. **(Afzal et al. 2019)**

Single sample network analysis

We began by building the aggregate reference networks for ssPCC and LIONESS: We used the LIONESS-S approach here as referenced in section *LIONESS: Linear Interpolation to Obtain Network Estimates for Single Samples* as this is the approach put forth in the original publication **(Kuijjer, Tung, et al. 2019)**. The reference networks are given in Figure 2.8A and Figure 2.9A, respectively.

As expected the two aggregate results/reference networks are rather different, with different relevant patterns of correlations. We recall that in this case, for ssPCC the reference network is the correlation matrix obtained from the control group, while for LIONESS the aggregate is obtained from the correlation of the complete data set (NSTI + control samples).

As it can be seen there are obvious differences. While the ssPCC reference network seems to cluster around Maltose, the LIONESS aggregate network seems to cluster around Valine and Ribitol.

We performed pathway analysis on the top 25 perturbed edges for both networks with the aim of assessing network properties and identifying structural and functional units in the metabolic networks **(Klamt et al. 2003)**. We found significant enrichment for aminoacyl-tRNA biosynthesis (P-value = 3.29×10^{-5}, fdr = 0.0028) and lysine degradation pathways (P-value = 1.33×10^{-4}, fdr = 0.0056) for the LIONESS aggregate network (Figure 2.8A) and significant enrichment for the aminoacyl-tRNA biosynthesis (P-value = 2.45×10^{-5}, fdr = 0.002) and valine, leucine and isoleucine biosynthesis pathways (P-value = 1.15×10^{-4}, fdr = 0.005) for the ssPCC reference network (Figure 2.9A). These results indicate that the single sample networks are

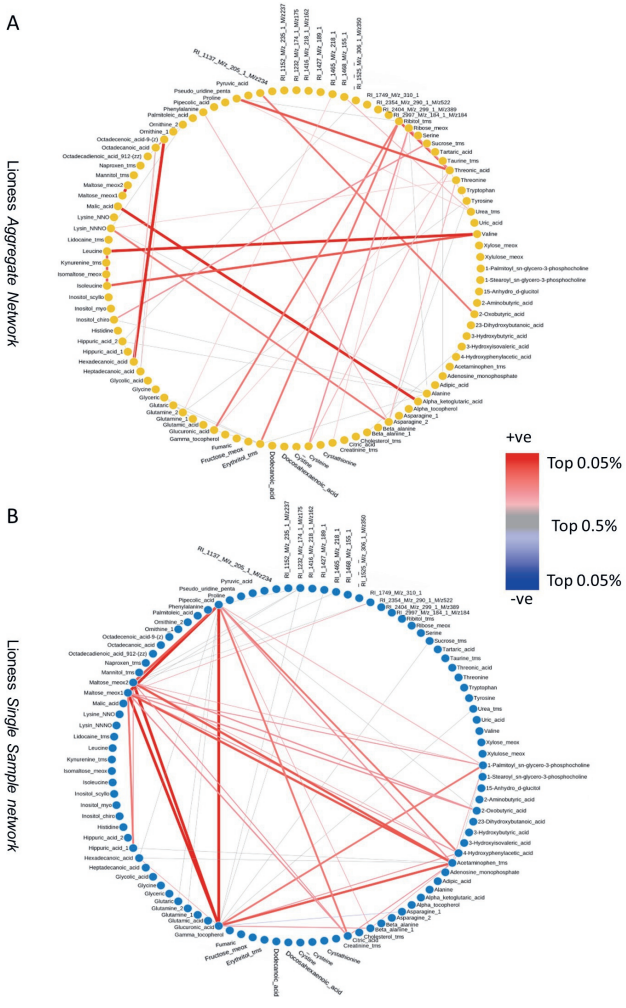


Figure 2.8: A) Aggregate (reference) network for Lioness B) Lioness single sample network for NSTI patient n. 24. The top (in absolute value) 0.05% edges are shown. The link width is proportional to the edge weight.

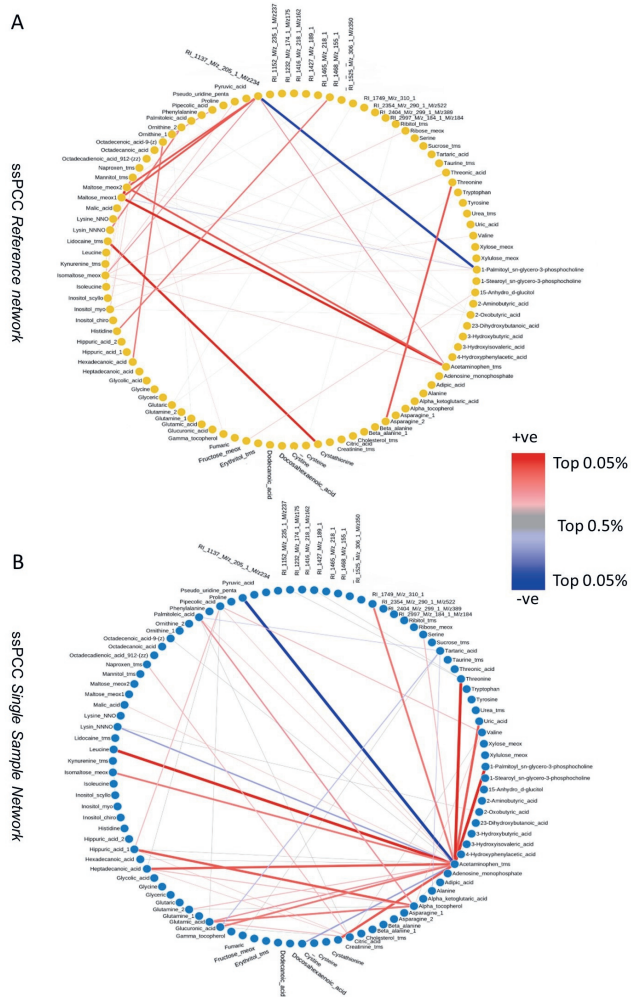


Figure 2.9: A) Aggregate (reference) network for ssPCC B) ssPCC single sample network for NSTI patient n. 24. The top (in absolute value) 0.05% edges are shown. The link width is proportional to the edge weight.

going to be constructed against a background that may encode for different biological phenomena and as such they will bear different biological information. This can be seen also from the results shown in Figure 2.5 where ssPCC and LIONESS edges are contrasted: two different reference networks result in related but different single sample edges.

It is interesting to note that no significant enrichment (after adjustment for multiple corrections) was found when pathway enrichment analysis was performed only on differentially abundant metabolites (see Table 3 in Afzal et al. **(Afzal et al. 2019)**): this is a clear indication that information on metabolic rewiring and/or disruption is reflected not only in change in metabolite levels but also in changes in the correlations between metabolite concentrations. In this case, single sample edges carry more information about changes in metabolism in NSTI than simple metabolite abundances.

We constructed the single sample networks for the 34 NSTI patients: as shown in Figure 2.10 the single sample networks obtained using ssPCC and LIONESS are markedly different, confirming what was observed using simulated data (see Section *Data simulations* and Figures 2.6 and 2.7): LIONESS and ssPCC networks cluster separately and, in general, LIONESS sample edges show higher variability than the corresponding ssPCC edges. In particular, there is a group of networks, corresponding to samples 5, 6, 10, 15, 20, 28, and 29) that are markedly different from the others: this is particularly evident for the LIONESS edges. All samples belong to patients with concurrent comorbidities with NSTI; all these patients are female, except patients 28 and 29.

Moreover, the ssPCC edges of sample 18 are more similar to LIONESS edges than to the other ssPCC samples.

We then focused on single sample networks built from the same sample profile using the two methods: we built correlation matrices among ssPCC and LIONESS edges and selected the NSTI patient for which the single sample networks obtained with ssPCC and LIONESS were most different (*i.e* the least correlated). We recovered from this analysis the single sample network for sample 24 which is a NSTI patient having a polymicrobial etiology. The corresponding single sample networks for ssPCC and LIONESS are given in Figure 2.8B and 2.9B where only the largest edges are shown (see Figure caption for more details).

As it can be seen there are obvious differences. In the LIONESS single sample network (Figure 2.8B) the edges connecting acetaminophen and glucuronic acid are the most disrupted. In the ssPCC single sample network (Figure 2.9B), the edges connecting acetaminophen, α-tocopherol (vitamine E), maltose and proline are altered. In the original publication **(Afzal et al. 2019)**, standard differential network analysis was performed to compare metabolite-metabolite connectivity in NSTI and surgery control (see Table 4 in Afzal *et al*. **(Afzal et al. 2019)**), glucuronic acid, and maltose were among the most differentially connected metabolites but not acetaminophen and α-tocopherol.

Univariate analysis of the single sample network-edges

Further, we compared the edges of LIONESS and ssPCC single sample networks between NSTI patients and controls using a t-test and we compared the results with

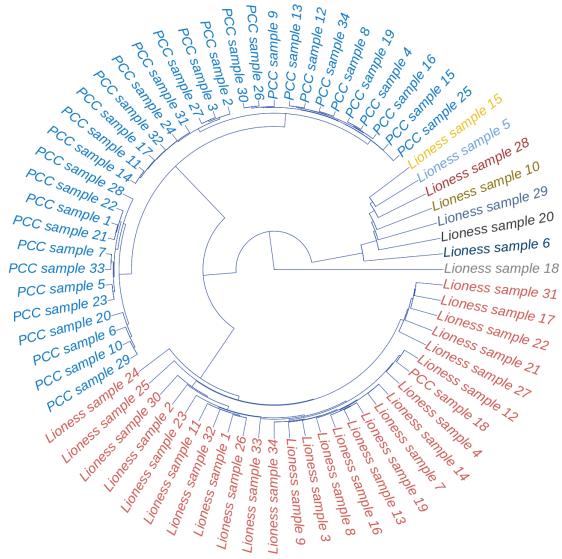


Figure 2.10: Clustering dendogram of the single sample networks for the metabolomics study case. The single samples for the subjects affected by necrotizing soft tissue infections are shown. Samples are colour coded by distance/similarity: red, for LIONESS (using LIONESS-S implementation) and blue for ssPCC single sample networks. Distance are calculated on the vectorized networks.

those obtained in the original publication using the metabolite concentrations. In particular, we compared NSTI and the control group on :

1. Metabolite concentrations
2. Metabolite-metabolite single sample network edges defined using ssPCC
3. Metabolite-metabolite single sample network edges defined using LIONESS-S
4. Metabolite-metabolite single sample network edges defined using LIONESS-D

Results are given in Table 2.4 which contains the top 10 metabolites/edges (sorted by P-value). In general, we observed that metabolites whose concentration is different between NSTI patient and controls are in general those whose single sample edges are found to discriminate among the two groups.

Random forest analysis on single sample network edges

Following the same strategy outlined in Section *Experimental data*, we performed Random Forest classifications to analyse the prediction accuracy of single sample edges in comparison with the simple correlations. In particular, Random Forest classification models were built for the following two-group comparisons:

Table 2.4: Results of *t*-test on the single sample edges obtained using Lioness (two implementations, mono and double) and ssPCC together with results from a *t*-test on the metabolite concentrations (column "Conc"). The top 10 edges (sorted by *P*-values) are given.

	Concentrations	ssPCC edges	Lioness-S edges	Lioness-D edges
1	1-stearoyl sn-glycro-3-phosphocholine	RI 2997 M/z 184 1 M/z184 & RI 2354 M/z 290 1 M/z522	Cystathionine & Glutaric acid	RI 2997 M/z 184 1 M/z184 & RI 2354 M/z 290 1 M/z522
2	1-palmitoyl sn-glycero-3-phosphocholine	RI 2997 M/z 184 1 M/z184 & RI 1232 M/z 174 1 M/z175	Cystathionine & Naproxen tms	RI 2997 M/z 184 1 M/z184 & RI 1232 M/z 174 1 M/z175
3	isomaltose meox	Alpha tocopherol & RI 2997 M/z 184 1 M/z184	Lidocaine tms & Glutaric acid	Alpha tocopherol & RI 2997 M/z 184 1 M/z184
4	RI 1416 m/z 218 1 m/z 162	Alpha tocopherol & Tryptophan	Cystathionine & Lidocaine tms	*a*-tocopherol & Tryptophan
5	*a*-tocopherol	1-Palmitoyl sn-glycero-3-phosphocholine & Maltose meox2	Pseudo uridine penta & Threonic acid	1-Palmitoyl sn-glycero-3-phosphocholine & Maltose meox2
6	tryptophan	Isomaltose meox & RI 1232 M/z 174 1 M/z175	3-Hydroxyisovaleric acid & Beta alanine	Isomaltose meox & RI 1232 M/z 174 1 M/z175
7	ribose meox	1-Palmitoyl sn-glycero-3-phosphocholine & Maltose meox1	2,3-Dihydroxybutanoic acid & Glycine	1-Palmitoyl sn-glycero-3-phosphocholine & Maltose meox1
8	glucoronic acid	RI 2997 M/z 184 1 M/z184 & Fumaric acid	Heptadecanoic acid & Asparagine	RI 2997 M/z 184 1 M/z184 & Fumaric acid
9	RI 1427 m/z 189 1	Citric acid & Glutamine-2	Octadecanoic acid & Asparagine	Citric acid & Glutamine-2
10	RI 2354 m/z 290 1 m/z 522	Alpha tocopherol & RI 1416 M/z 218 1 M/z162	Octadecenoic acid-9-(z) & Asparagine	Alpha tocopherol & RI 1416 M/z 218 1 M/z162

Table 2.5: Accuracy of the Random Forest models constructed using the concentration/abundance profiles and the edges of single sample networks built from the metabolomics data set investigating necrotising soft tissue infections

Model	Classification accuracy			
	Conc	ssPCC	LIONESS-S	LIONESS-D
NSTI *vs* Controls	94.7	98.3	98.3	100.0
Mono *vs* Poly	87.1	82.5	82.5	82.5
Streptococcus vs Staphylococcus	85.6	75.9	75.9	79.3

1. NSTI (n = 34) vs Controls (n = 24)
2. Mono-microbial NSTI (n = 26) vs poly-microbial NSTI (n = 7)
3. *Streptococcus* NSTI (n = 20) vs *Staphylococcus aureus* (n = 8)

The models were built using the original concentration/abundance profiles, the edges of the single sample networks built using ssPCC, the edges of the single sample networks built using LIONESS-S, and the edges of the single sample networks built using LIONESS-D. Classification accuracies are given in Table 2.5.

Regarding the comparison between NSTI and controls, the use of the single sample network edges increases the accuracy (up to 100%), although the use of simple concentrations gives excellent classification (94.7%). The logic behind the use of single sample network edges is that additional information is contained in the relationships (or disruption thereof) among (pairs of) metabolites that are contained in, or is additional to, the metabolite levels and thus better accuracy should be, in principle obtained. However, this is not always the case: for the comparison between mono versus poly microbial infection and *Streptococcus* versus *Staphylococcus* infection the use of single sample network results in decreased accuracy. This may well depend

Table 2.6: Random Forest classification of Necrotizing soft tissue infection patients and controls using metabolite concentrations, ssPCC and LIONESS single-sample network edges. The top 10 metabolites and metabolite-metabolite edges are shown in decreasing order of importance (given by the Mean Decrease Gini Index).

	Concentration	ssPCC edges	Lioness-S edges	Lioness-D edges
1	1-Palmitoyl sn-glycero-3-phosphocholine	1-Stearoyl sn-glycero-3-phosphocholine & 1-Palmitoyl sn-glycero-3-phosphocholine	Lidocaine tms & Glutaric acid	Alpha tocopherol & RI 2997 M/z 184 1 M/z184
2	1-Stearoyl sn-glycero-3-phosphocholine	RI 2997 M/z 184 1 M/z184 & RI 2354 M/z 290 1 M/z522	Naproxen tms & Lidocaine tms	RI 2997 M/z 184 1 M/z184 & Maltose meox1
3	RI 2997 M/z 184 1 M/z184	1-Palmitoyl sn-glycero-3-phosphocholine & RI 2354 M/z 290 1 M/z522	Pseudo uridine penta & Threonic.acid	RI 2997 M/z 184 1 M/z184 & RI 2354 M/z 290 1 M/z522
4	RI 1416 M/z 218 1 M/z162	1-Palmitoyl sn-glycero-3-phosphocholine & Citric acid	Cystathionine & Lidocaine tms	RI 2354 M/z 290 1 M/z522 & Tryptophan
5	Isomaltose meox	1-Stearoyl sn-glycero-3-phosphocholine & RI 2354 M/z 290 1 M/z522	Naproxen tms & Glycine	1-Stearoyl sn-glycero-3-phosphocholine & 1-Palmitoyl sn-glycero-3-phosphocholine
6	Ribose meox	1-Palmitoyl sn-glycero-3-phosphocholine & RI 2997 M/z 184 1 M/z184	Naproxen tms & Glutamine	1-Palmitoyl sn-glycero-3-phosphocholine & RI 2354 M/z 290 1 M/z522
7	Citric acid	1-Stearoyl sn-glycero-3-phosphocholine & Citric acid	Cystathionine & Glutaric acid	RI 2354 M/z 290 1 M/z522 & RI 1416 M/z 218 1 M/z162
8	Tryptophan	Alpha tocopherol & Tryptophan	Naproxen tms & Glutaric acid	RI 2354 M/z 290 1 M/z522 & Citric acid
9	Alpha tocopherol	1-Stearoyl sn-glycero-3-phosphocholine & Tryptophan	Octadecenoic acid-9-(z) & Asparagine-2	Alpha tocopherol – Tryptophan
10	RI 2354 M/z 290 1 M/z522	Tryptophan & RI 1416 M/z 218 1 M/z162	Naproxen tms & Lysine	RI 1416 M/z 218 1 M/z162 & Glyceric acid

on the limited sample size used to build the aggregated/reference networks that can lead to instability in the estimation of these networks **(Suarez-Diez et al. 2015)** and, as a consequence, low-quality estimation of the single sample networks. We also compared the edge and metabolite importance in the Random Forest models using the Mean Decrease Gini index as the importance measure. The top metabolites and edges from the models are shown in Table 2.6. The metabolites whose edges mostly contribute to the separation between NSTI and controls tend to be the metabolites whose concentration is also different between the two groups, confirming that the single sample edges bear content of both concentration and pairwise relationships as also discussed in Section *ssPCC: Single sample network based on Pearson's correlation*.

2.4 Conclusions

In this study we investigated and assessed the utility of two methods for the inference of single sample networks in a metabolomics context: LIONESS (Linear Interpolation to Obtain Network Estimates for Single Samples) **(Kuijjer, Tung, et al. 2019)** and ssPCC (single sample network based on Pearson's correlation)**(X. Liu, Y. Wang, et al. 2016)**.

The two methods are functionally related and when compared on the simulated data with different correlative properties, we found both methods to have limited ability to describe different situations. However on data from a NF-kB dynamic metabolic model we found only the ssPCC single sample networks are able to describe different situations arising from perturbations of the model.

We found the statistical procedure proposed in ssPCC to have limited power and to be heavily dependent on the particular reference networks and of little utility for practical applications.

We explore the potential of single sample edges to be used in place of concentration to discriminate between groups in a case-control scenario. To this scope we used two different implementations of LIONESS and proposed a work-around to adapt ssPCC to this scenario. Using Random Forest as a classification algorithm, we found that in 13 cases out of 22 the use of ssPCC edges gives better (or equal) classification accuracy than (to) the use of actual metabolite concentrations, while, overall, the use of LIONESS sample resulted in worse prediction models.

We found that single sample networks built off of a control group (like ssPCC and LIONESS-S) yield better results than those built off of similar samples in a classification setting, however this approach does not allow for generalization and as such should be used as an exploratory tool.

We finally applied the two methods to analyse a metabolomics study pertaining to necrotising soft tissue infections.

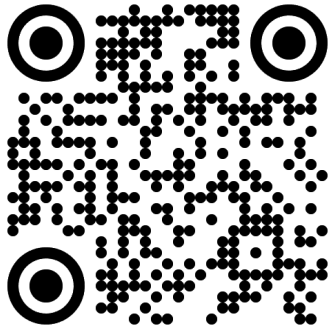
NSTI specific single sample networks obtained with the two methods are markedly different and are likely to describe different ongoing biological processes. We found that single sample edges, either from ssPCC or LIONESS, gave better prediction results in distinguishing between NSTI patient and controls but not in other comparisons aimed to distinguish between disease aetiology. In general, ssPCC edges found to be important to discriminate groups involves metabolite pairs that are found important when comparing groups with a standard *t*-test performed on concentrations.

2.5 Author contributions

E.S. designed the study and supervised the work. S.J. and E.S. performed data analysis. S.J and E.S. wrote the manuscript. All authors read and approved the final version of the paper.

2.6 Acknowledgements

This study has received funding from The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (research projects on personalized medicine-smart combination of pre-clinical and clinical research with data and ICT solutions).

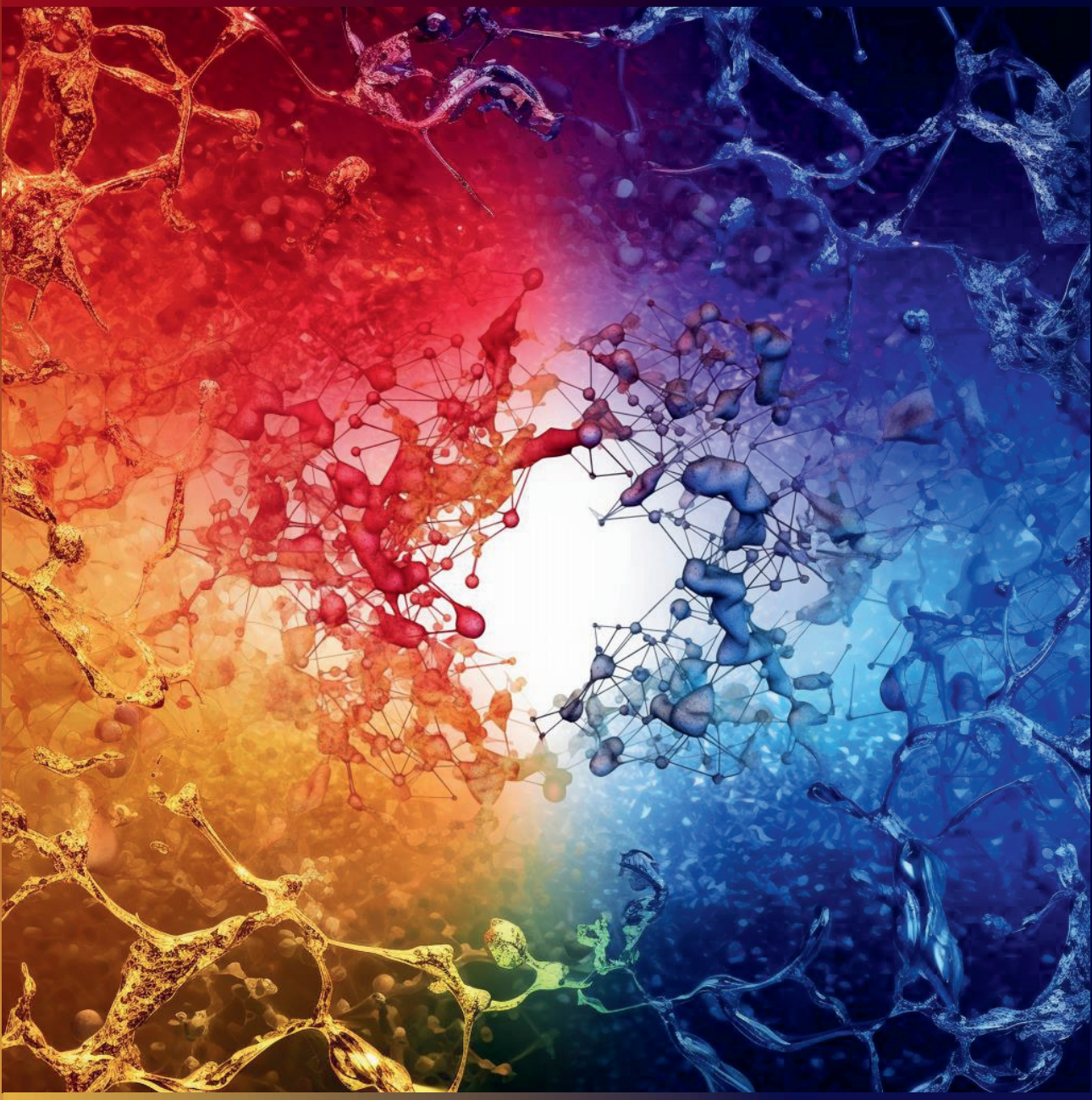


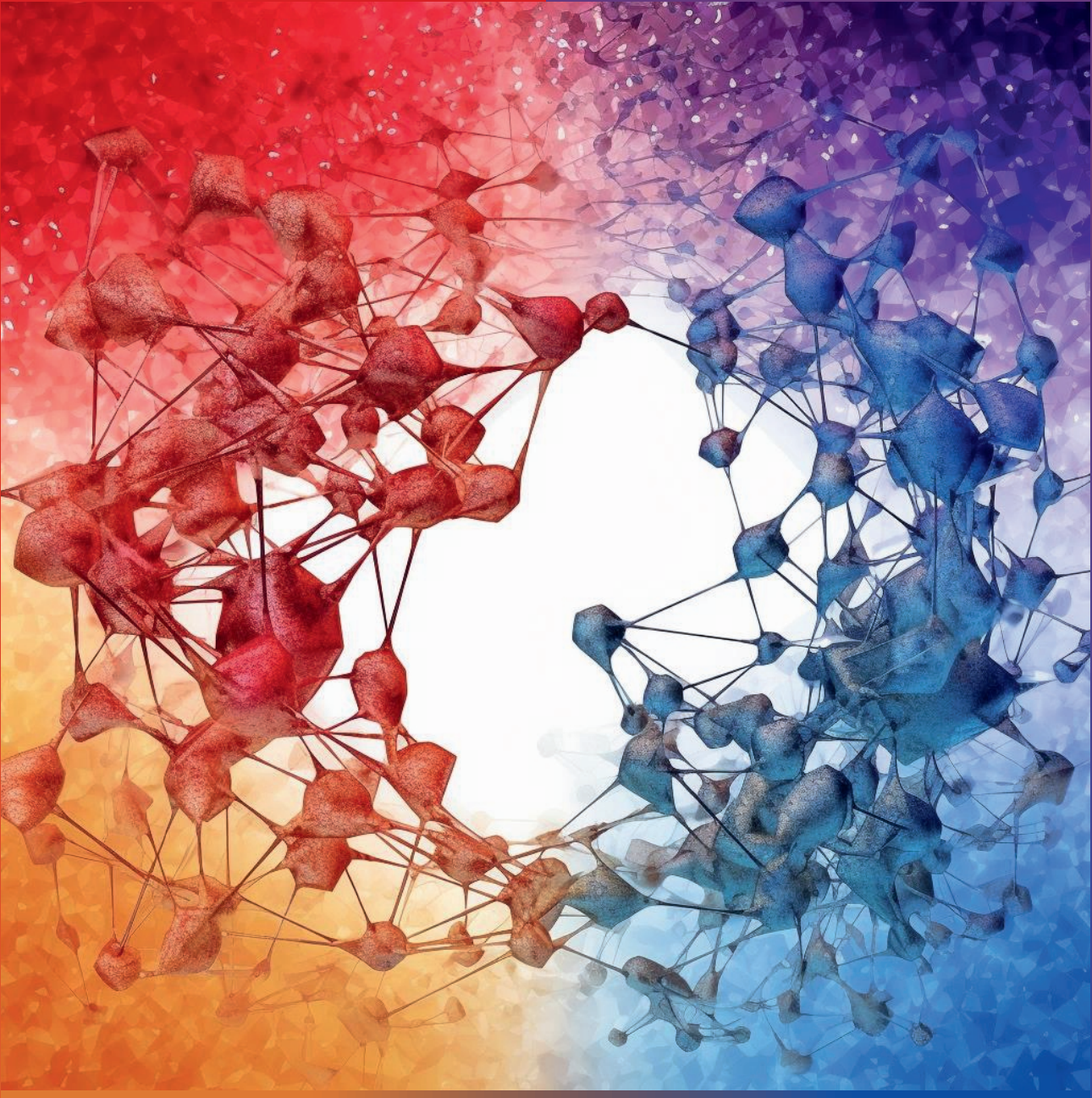
(a) INFECT, PerMIT & PerAID website

(b) Publication

Chapter 3

Chapter





Sanjeevan Jahagirdar[1], Edoardo Saccenti[1]

Turn to page 377 for author affiliations

This chapter is adapted from:
Jahagirdar, S., & Saccenti, E. (2020). On the Use of Correlation and MI as a Measure of Metabolite—Metabolite Association for Network Differential Connectivity Analysis. *Metabolites*, *10*(4), 171. https://doi.org/10.3390/metabo10040171

Correlation or Mutual Information? That is the Question!

Abstract

Metabolite differential connectivity analysis has been successful in investigating potential molecular mechanisms underlying different conditions in biological systems. Correlation and Mutual Information are two of the most common measures to quantify association for building metabolite-metabolite association networks and to calculate differential connectivity. In this study, we investigated the performance of correlation and Mutual Information to identify significant differentially connected metabolites. These association measures were compared on *i*) 23 publicly available metabolomic data sets and 7 data sets from other fields, *ii*) simulated data with known correlation structures, and *iii*) data generated using a dynamic metabolic model to simulate real-life observed metabolite concentration profiles. In all cases, we found more differentially connected metabolites when using correlation indices as a measure for association than Mutual Information. We also observed that different Mutual Information estimation algorithms resulted in difference in performance when applied to data generated using a dynamic model. We concluded that there is no significant benefit in using Mutual Information as a replacement for standard Pearson's or Spearman's correlation when the application is to quantify and detect differentially connected metabolites.

3.1 Introduction

Metabolite concentration profiles measured in samples like blood, urine or tissues, and their patterns of variations, are regulated by complex bio-molecular machines. In recent times, there has been a shift towards studying metabolite profiles in a holistic manner by computational and mathematical methods thanks to the possibility of measuring many metabolites simultaneously using high-throughput techniques like Mass spectroscopy (MS) and Nuclear magnetic resonance (NMR) **(Tavassoly et al. 2018; Vignoli, Ghini, et al. 2019; Emwas et al. 2019)**.

A biological system can be represented as a complex network of interconnected biomolecular entities **(Ma'ayan 2011)** which can be visualised in a graphical manner as networks, *i.e.* sets of nodes that are connected by edges to indicate the existence and the strength of pairwise relationships **(Trudeau 2013)**. This representation shifts the focus towards the relationships among biological entities rather than on their levels; in this light, network and network analysis are fundamental tools from the systems biology toolbox to investigate and understand metabolomic data **(Rosato, Tenori, Cascante, Carulla, et al. 2018)**. When the nodes are metabolites, the network can be called metabolite-metabolite association network **(Saccenti, Suarez-Diez, et al. 2015; Rosato, Tenori, Cascante, Carulla, et al. 2018)**, and in modern metabolomic studies, the interest is to reconstruct these association patterns from observed data measured in well designed experiments.

Association patterns are usually quantified using similarity measures like correlation and Mutual Information (MI) and most algorithms built for the purpose of network inference make use of one of these two indices **(Jahagirdar, Suarez-Diez, et al. 2019)**.

Once metabolite-metabolite association networks are reconstructed, they can be analysed in the context of the study design they have been reconstructed, for instance comparing them across two or more conditions and performing a so-called differential network analysis. In particular, the interest lies in comparing the connections and magnitude thereof for each metabolite between different networks to highlight network differences. The rationale is that under normal conditions of the system the metabolites behave in an orchestrated manner and perturbations to the systems, such as those induced by pathophysiological conditions will induce modifications in the relationships among metabolites, which will be reflected in their connectivity patterns. Metabolite connectivity and differential connectivity analysis are illustrated in Figure 3.1.

In metabolomics, metabolite differential connectivity analysis has been successful to investigate and highlight potential molecular mechanisms underlying cardiovascular diseases **(Saccenti, Suarez-Diez, et al. 2015)**, age and sex phenotypes **(Vignoli, Tenori, Luchinat, et al. 2017)**, acute myocardial events **(Vignoli, Tenori, Giusti, et al. 2020)** and severe bacterial infections **(Afzal et al. 2019)**. For instance, Saccenti *et al.* **(Saccenti, Suarez-Diez, et al. 2015)** analysed the metabolite-metabolite association networks specific to different cardiovascular risk patients and reported differential connectivity of Very Low Density Lipoprotein (VLDL) and glucose in high and low risk networks. Azal *et al.* **(Afzal et al. 2019)** found the networks specific to patients with necrotising soft tissues infections to be more connected than those of healthy

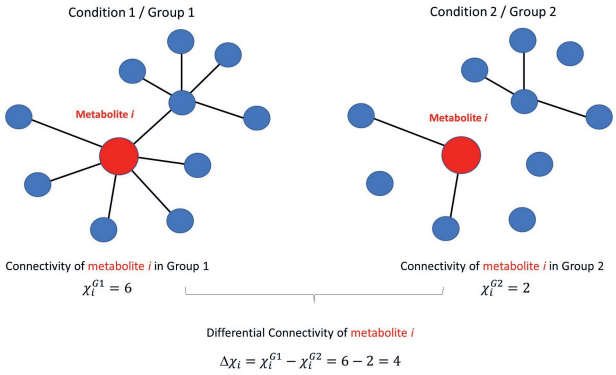


Figure 3.1: Graphical illustration of the concept of metabolite connectivity and differential connectivity. An ideal unweighted metabolite-metabolite association network involving 10 metabolites is shown under two different conditions. Metabolite i is connected with a different number of metabolites (represented by the existence of an edge) in the two conditions. The connectivity χ_i of metabolite i is given by the number of connecting edges (for a generalisation for weighted association networks, see Equation 3.23 in Section 3.3.3): 6 under condition 1 and 2 under condition 2. The differential connectivity of metabolite i is given by $\Delta\chi_i = \chi_i^{G1} - \chi_i^{G2} = 6 - 2 = 4$ as described in Equation (3.24).

controls and singled out differentially connected metabolites that showed capability of interfering with bacterial biofilm formation.

Motivation for this study arose when re-analysing data from **(Rist et al. 2017)** in the context of differential analysis of metabolite-metabolite association networks. The original study dealt with the characterisation of metabolites profile associated with sex and age; we were interested in exploring sex-specific patterns of metabolite-metabolite association networks. To this aim we performed differential network analysis as detailed in the *Methods* section; briefly metabolite-metabolite association networks where built starting from the sample correlation matrices or the Mutual Information calculated from male and female samples and a weighted connectivity was calculated as the sum of the (absolute) values of the pairwise Pearson's correlation (respectively, Mutual Information) of a metabolite with every other metabolites, as illustrated in Figure 3.2. Differential connectivity was defined as the difference between each metabolite connectivity in male and female specific networks, as exemplified in Figure 3.1. Significance was assessed using a permutation test.

We observed many more differentially connected metabolites when using correlations as a measure of association than with Mutual Information. Actually, all 128 measured metabolites showed statistically significant differential connectivity when correlation was used and only 23 when Mutual Information was used. These results were at first surprising: we expected Mutual Information to be a more informative measure for quantifying relationship among metabolite than Pearson's correlations. After all, it is a common place to expect metabolites to exhibit non-linear behaviour which is better captured by Mutual Information. Mutual Information (see definitions and Equations in section 3.2.3) is a non parametric measure and it is a comprehensive measure of independence which makes it superior (in principle) for accounting for

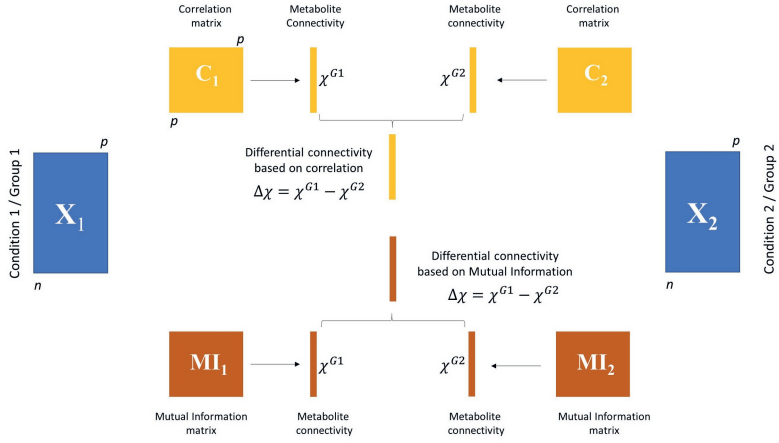


Figure 3.2: Graphical illustration of differential connectivity analysis. Given two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ of size $n_1 \times p$ and $n_2 \times p$ with n_1 possibly different from n_2, weighted association matrices are built using either correlation ($\mathbf{C}_1$ and $\mathbf{C}_2$, for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively) or Mutual Information ($\mathbf{MI}_1$ and $\mathbf{MI}_2$). Weighted (metabolite) connectivity is then calculated as described in Equation 3.23 for group 1 and group 2 as χ_i^{G1} and χ_i^{G2}. The differential connectivity is given by $\Delta\chi_i = \chi_i^{G1} - \chi_i^{G2}$ and it is calculated using both correlation and Mutual Information. Significance is then assessed using a permutation test.

both linear and nonlinear dependencies **(R. Smith 2015)**. In fact Pearson's correlation can underestimate the dependence between variables when the dependence translate into non-linear relationships.

An illustrative example is given in Figure 3.3 that shows four different data patterns (plot of simulated metabolite concentration) all having the same Mutual Information (1.32 nats) but very different correlation. Correlation is not able to capture highly non-linear dependence like in the case shown in panel C, where the metabolites are obviously interdependent.

The question arose of why we observed such counter intuitive behaviour which led us to explore the question of which association measure is more appropriate for differential analysis of metabolite-metabolite association networks. We started by re-analysing 23 data sets of publicly available metabolomics studies from several research fields, ranging from plant to cancer metabolomics, acquired on different matrices, from cell to tissues, with both MS and NMR. We then compared Mutual Information and correlation on simulated data with different correlation structures and properties and using different algorithms to estimate Mutual Information (see Results section *Differential connectivity analysis on experimental data*). Finally we also compared Mutual Information and correlation on simulated data generated using a dynamic NF-$\kappa\beta$ metabolic model. In all cases we found correlation, either of Pearson's or Spearman's formulation, to be a more sensitive measure of similarity than Mutual Information when used in the context of differential connectivity analysis.

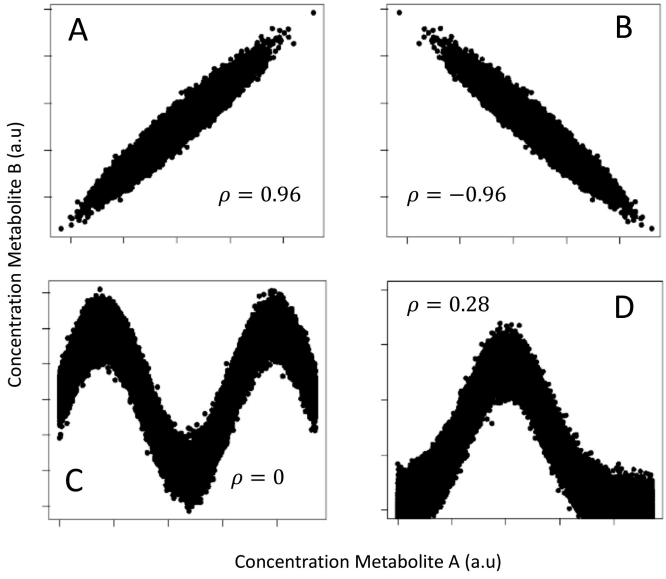


Figure 3.3: Four different data pattern obtained by plotting the simulated concentration of two metabolites A and B on which Gaussian experimental noise has been added. A) Positive linear relationship, $\rho = 0.96$ (Pearson's correlation); B) Negative linear relationship, $\rho = -0.96$; C) Sine-wave relationship, $\rho = 0$; D) Bell-shaped relationship, $\rho = 0.28$. In all cases the Mutual Information is 1.32 nats (or 1.90 bits). One nat is the information content of the uniform distribution on the interval $[0, e]$ where e is the basis of the natural logarithm. This figure is an adaptation from Table 1 from reference **(R. Smith 2015)**.

3.2 Methods

3.2.1 Association measures

In this study we use two methods to calculate correlations and four methods to estimate Mutual Information as association measures for building the networks

3.2.2 Correlation indices

Pearson and Spearman correlation measures

The Pearson's (sample) correlation coefficient **(Pearson 1895)** between two random variables X and Y is defined as

$$\rho = \frac{\text{cov}(X, Y)}{S_X \times S_Y} \,, \tag{3.1}$$

where S_X and S_Y is the standard deviation of the measured X variables (respectively, Y) and $\text{cov}(X, Y)$ is the covariance between X and Y. The Pearson correlation coefficient is probably the most used measure of association used in life sciences and it is a standardised version of the covariance, which being dependent on the scale of

the variables can vary, in principle, between 0 and $+\infty$. The Spearman's correlation coefficient **(Spearman 1904)** between two variables X and Y is defined as

$$\rho_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)} , \tag{3.2}$$

where d is the difference in rank order between metabolite X and Y and n is the sample size. The Spearman's correlation coefficient is an appropriate measure for non linear association between two variables X and Y.

3.2.3 Mutual Information

Mutual Information is defined in information theory as the mutual dependence of two random variables X and Y and can be interpreted as reduction in uncertainty of the outcome of one variable on observation of another variable. Before defining operatively the concept of Mutual Information, we shall introduce the concept of entropy since it is related to MI. Entropy is a measure of the uncertainty about the values that a certain random variable X, distributed with probability distribution $p(x)$, can assume.

$$H(X) = -\sum p(x)\log p(x) , \tag{3.3}$$

while if X is continuous

$$H(X) = -\int p(x)\log p(x)dx , \tag{3.4}$$

Equation (3.3) can be recognised as the the expectation value of $-\log p(x)$, thus

$$H(X) = \mathrm{E}[-\log p(x)] . \tag{3.5}$$

As an example, assuming a metabolite X whose concentration can assume only the values $x_1 = 0.4$, $x_2 = 0.9$ and $x_3 = 1.3$ with probability $p(X = x_1) = 0.2$, $p(X = x_2) = 0.7$ and $p(X = x_3) = 0.1$, the entropy of X is

$$\begin{aligned} H(X) &= -\sum_{x_1,x_2.x_3} p(x)\log p(x) && (3.6)\\ &= -[0.4\times\log(0.4) + 0.7\times\log(0.7) + 0.1\times\log(0.1)] && (3.7)\\ &= 0.8018 . && (3.8) \end{aligned}$$

The entropy measures the uncertainty of a variable: the higher the entropy, the higher the uncertainty on that variable. Turning to a biological example, if a metabolite shows little variability, *i.e.* its range of variation is limited, its entropy will be also lower. On the contrary, a metabolite with a large variability will have high entropy. The entropy is usually related to the content of information of a random variable: the higher the entropy, the higher the information content. One can think of a metabolite that does not vary, whatever the experimental circumstances, that is it assumes value c with probability $p(X = c) = 1$ its entropy will be $H(X) = 0$ and thus null the information associated to it.

Thus, the calculated entropy of a metabolite will be related to its variance. For instance, if X is normally distributed $\approx N(\mu, sigma^2)$, its entropy is just $\frac{1}{2}\left(\log 2\pi\sigma^2 + 1\right)$. The entropy of a variable is maximum when its probability distribution is uniform and, in contrast with the variance, it can assume negative values.

In practical applications the probability distribution $p(x)$ is not known *a priori*, but is estimated from the observed distribution of the data, *i.e.* the empirical entropy is estimated. Estimating entropy is not a trivial task and many different algorithms exists.

The most common way of expressing the Mutual Information between two random variables X and Y is by expressing the distance between the joint distribution $p(X,Y)$ and product distribution $p(X)p(Y)$ using the Kullback-Leibler divergence **(Kullback et al. 1951)**:

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{3.9}$$

Since

$$\log\Big(\frac{p(x,y)}{p(x)p(y)}\Big) = log\Big(\frac{p(x|y)}{p(x)}\Big) = log\Big(\frac{p(y|x)}{p(y)}\Big), \tag{3.10}$$

it follows that

$$MI(X,Y) = H(X) - H(X|Y), \tag{3.11}$$

and taking into account the symmetry of information

$$H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{3.12}$$

an elegant expression of Mutual Information as a function of Entropy where the Mutual Information $MI(X,Y)$ between X and Y can be obtained as

$$MI(X,Y) = H(X) + H(Y) - H(X,Y), \tag{3.13}$$

where $H(X)$ and $H(Y)$ is the entropy of X and Y, respectively and $H(X,Y)$ is the entropy of X and Y. Hence, the problem of estimating Mutual Information boiled down to the problem of estimating entropy. In this study we use four different methods to estimate entropy in order to calculate Mutual Information as implemented in the `infotheo` R package **(P. E. Meyer 2008)**.

Entropy of empirical probability distribution

The most common approach to estimate entropy is through the calculation of the probability distribution starting from the empirical data **(P. E. Meyer 2008)**. This is obtained by computing the relative frequency of occurrence of each value:

$$\hat{H}^{emp}(X) = -\sum_{x\in X}\frac{\#(x)}{n}\log\frac{\#(x)}{n}, \tag{3.14}$$

where $\#(x)$ is the number of data points having value x and n is the number of samples. However, it is necessary to note that empirical estimators are biased downwards and the estimate is always smaller than actual entropy and the variance of the empirical estimator is dependent on the sample size **(Paninski 2003)**. More precisely, the variance is upper bounded by $\left(\frac{(log\ n)^2}{n}\right)$.

Miller-Madow asymptotic bias corrected empirical estimator

The empirical estimation suffers from an asymptotic bias of $-\frac{|x|-1}{2n}$ where $|x|$ is the number of bins with non-zero probability. This bias can be especially large if the number of bins start exceeding the sample size. The Miller-Madow correction attempts to get around this problem by adding the asymptotic bias to the empirical estimation of entropy **(Paninski 2003)**. This correction is given by

$$\hat{H}^{mm}(X) = \hat{H}^{emp}(X) + \frac{|x|-1}{2n} , \tag{3.15}$$

and it reduces the bias of the estimation without changing the variance.

Shrinkage estimate of the entropy of a Dirichlet probability distribution

Shrinkage is a popular technique to improve estimators, especially for smaller sample sizes. The shrinkage estimator attempts to combine two estimators in a weighted average with a factor $\lambda^* \in [0,1]$. The two estimators are as follows,

$$\frac{1}{|x|} , \tag{3.16}$$

$$\frac{\#(x)}{n} . \tag{3.17}$$

The method shrinks the latter estimate towards the former by minimising the mean square error λ^*. The entropy estimate is then given by

$$\hat{H}^{shrink}(X) = -\sum_{x \in X} \hat{p}\lambda^*(x) \log \hat{p}\lambda^*(x) , \tag{3.18}$$

where

$$\hat{p}\lambda^*(x) = \lambda^* \frac{1}{|x|} + (1-\lambda^*)\frac{\#(x)}{n} . \tag{3.19}$$

The target estimator $\frac{1}{|x|}$ has low variance and high bias whereas the unregulated estimator $\frac{\#(x)}{n}$ has large variance and low bias. The benefit of using such a shrinkage method is that the resulting estimator surpasses both of the individual estimates in terms of accuracy and statistical efficiency **(J. Schäfer and Strimmer 2005b; J. Schäfer and Strimmer 2005a)**.

Schurmann-Grassberger estimation

The Schurmann-Grassberger method estimates the entropy by utilising a Bayesian parametric strategy assuming samples to be Dirichlet distributed, *i.e.* multivariate beta distributed given by

$$p(X;\theta) = \frac{\prod_{i \in \{1,2,\ldots|x|\}} \Gamma(\theta_i)}{\Gamma(\sum_{i \in \{1,2,\ldots.|x|\}} \theta_i)} \prod_{i \in \{1,2,\ldots.|x|\}} x_i^{\theta_i - 1} . \tag{3.20}$$

The entropy of the Dirichlet distribution can be determined by the following with $\theta_i = N$ as a constant probability of every event.

$$\hat{H}^{dir}(X) = \frac{1}{n + |X|N} \sum_{x \in X} (\#(x) + N)(\psi(n + |X|N + 1) - \psi(\#(x) + N + 1)) \,, \tag{3.21}$$

where, N is the prior probability of an event $x_i \in X$ assuming that no event x_i becomes more probable than another, and $\psi(z)$ as the Digamma function with $\psi(z) = \frac{\mathrm{d} ln \Gamma(z)}{\mathrm{d}z}$ and $\Gamma(z)$ as the Gamma function **(Nemenman et al. 2004; Schürmann et al. 1996; Wu et al. 2007)**. It should be remarked that all the estimations used above assume the variables to be discrete in nature; continuous variables are binned before calculations as a pre-processing step. We used the default binning parameters from `infotheo` R package.

3.2.4 Network concepts

A network or graph is a graphical representation of the association between objects. In biology such are molecular components like genes, proteins, or metabolites and in the network they are represented by nodes. The association between two nodes is represented as link (or edge) connecting the two nodes. The nature of the association among the molecular features can be diverse: in the case of genes regulatory networks, the edges represent regulatory interactions where the protein product of a given gene directly modulates the expression of a target gene; in co-expression networks, the edge represents significant co-expression levels of the connected genes; in protein-protein interaction networks, edges represent the existence of a physical interaction between proteins. In metabolite-metabolite association networks, two metabolite are connected if their concentration levels are significantly correlated.

For manipulation and analysis, networks can be mathematically represented as matrices through the so-called adjacency (also called connectivity) matrix A: the rows and columns of the adjacency matrix represent the nodes whereas non-null entries represent links. If the edges are binary indicating only the presence-absence of an association the network is said to be *unweighted* and the elements a_{ij} of the adjacency matrix describing the association between node i and j are either 1 or 0:

$$a_{ij} = \begin{cases} 1 & \text{if there is association} \\ 0 & \text{otherwise} \end{cases} \,. \tag{3.22}$$

If the strength of the interaction can be quantified, a weight can be given to the edge and thus, the network is said to be *weighted*: in this case, the elements of a weighted adjacency matrix are real numbers that indicate the strength of the interaction and can vary, for instance, in the $[-1,1]$ range for correlation, in the $[0,+\infty]$ range for Mutual Information, or in the $[0,1]$ range for probability. Each node in a network can be characterised using functions that can be derived from the patterns of its association. A very common measure is the node degree or connectivity, that is the number of its connection. For a $p \times p$ network A, the connectivity of the node i is

given by

$$\chi_i = \sum_{j>i} |a_{ij}| \,. \tag{3.23}$$

If the network is unweighted, it holds $0 < \chi_i < p-1$. If the network is weighted, the range of the connectivity depends on the nature of the association measure. If (the absolute value of the) correlation is used, χ_i still ranges between 0 and $p-1$ in which case it means that the molecular feature represented by node a_i is perfectly correlated with all other nodes in the network. If Mutual Information is used, which is in the $[0, +\infty)$ range, χ_i also range between 0 and ∞.

Differential network analysis

Differential connectivity (see Figure 3.1 for a graphical overview) is calculated comparing the metabolite connectivity for p metabolites measured under two different conditions or in two groups, as exemplified in Figure 3.2.

Given two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ of size $n_1 \times p$ and $n_2 \times p$ with n_1 possibly different from n_2, measured under Group 1 (condition 1) and Group 2 (condition 2), respectively (total sample size $n = n_1 + n_2$), and selecting an association measure (either correlation or Mutual Information), the differential connectivity $\Delta\chi_i$ for the ith node (metabolite) is given by

$$\Delta\chi_i = \chi_i^{G1} - \chi_i^{G2} \,. \tag{3.24}$$

In the simulation study discussed in Section *Data simulations*, data $\mathbf{X}_2$ is taken to be $\approx N(0, \mathbf{I}_p)$ where $\mathbf{I}_p$ is the identity matrix of appropriate dimensions. Under this model the expected connectivity $\mathrm{E}[\chi_i^{G2}]$ (where $\mathrm{E}[*]$ indicate the expected value of $*$) is zero, from which it follows that

$$\mathrm{E}[\Delta\chi_i] = \mathrm{E}[\chi_i^{G2} - \chi_i^{G1}] = \mathrm{E}[\chi_i^{G1}] = \chi_i^{G1}. \tag{3.25}$$

Permutation tests to assess the statistical significance of differential connectivity

The significance of the differential connectivity was assessed implementing a permutation test. First, each and every column of the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ pertaining to Group 1 and 2 (see Figure 3.2) is independently permuted; the column values $x_1, x_2, \ldots x_n$ are replaced by $x_{p(1)}, x_{p(1)}, \ldots, x_{p(n)}$ where $p(1), p(2), \ldots, p(n)$ are random permutation of $1, 2, \ldots, n$. This ensures that the mean and the variance of each column in $\mathbf{X}_1$ and $\mathbf{X}_2$ are preserved, but the relationships among the variables are destroyed. For randomised data the expected metabolite connectivity is $\mathrm{E}[\chi_i] = 0$.

The permuted version of $\mathbf{X}_1$ and $\mathbf{X}_2$ are used to build the weighted association matrices, using either correlation or Mutual Information which are then used to compute, for each metabolite, the "permuted" differential connectivity:

$$\Delta\chi_i^{perm} = \chi_i^{G1,perm} - \chi_i^{G2,perm} \,. \tag{3.26}$$

The permutation procedure is repeated $N_{perm} = 10^3$ times to build a distribution D_i of permuted differential connectivity values for metabolite i. This distribution is which

is used to compute the significance of the differential connectivity of metabolite i, which is expressed as P-value calculated as

$$P_i = \frac{1 + \text{Num}(D_i > \Delta\chi_i)}{N_{perm}} , \tag{3.27}$$

where $\text{Num}(D_i > \Delta\chi_i)$ indicates the number of elements of D_i whose absolute value is larger than χ_i, the differential connectivity of metabolite i calculated from the original, non-permuted, data $\mathbf{X}_1$ and $\mathbf{X}_2$. This permutation approach is equivalent to a hypothesis testing procedure where the null hypothesis

$$\text{H}_0 : \Delta\chi_i = 0 , \tag{3.28}$$

is tested again the alternative hypothesis

$$\text{H}_1 : \Delta\chi_i > 0 . \tag{3.29}$$

3.2.5 Data simulations

Data were randomly generated under a Gaussian multivariate model with $\mathbf{X}$ a $n \times p$ data matrix

$$\mathbf{X} \approx N(\mathbf{0}, \Sigma_p) , \tag{3.30}$$

with n varying between 10 and 1000. All variables have been simulated with variance equal to 1, so Σ equals the correlation matrix. Three different correlation structures were used as described in the following section.

Toeplitz correlation structure

The Toeplitz correlation structure (also called the auto-regressive model) describes correlation patterns where adjacent pairs of observations are highly correlated, and those further away are less correlated, with the correlation between the i-th and j-th observations decay exponentially with respect to $|i - j|$.

This correlation structure is often used to simulate data in a linear discriminant setting **(Y. Guo et al. 2006)**, in linear mixed modelling, and in the time series literature as a model for group correlations **(Hardin et al. 2013)**.

The corresponding correlation matrix has the form

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{p-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{p-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \rho^{p-4} & \cdots & 1 \end{pmatrix} . \tag{3.31}$$

We generated 10 Toeplitz correlation matrix by varying ρ between 0.0 and 1.0 in steps of 0.1. Given ρ, random Toeplitz matrices were generated using the strategy proposed by Hardin and coworkers **(Hardin et al. 2013)** using the R function `simcorTop` provided in the supplementary material of **(Hardin et al. 2013)** and available at pages.pomona.edu/ jsh04747/research/simcor.r. The parameters used were $k = 1$, $\epsilon = 0.01$ and edim = 2. Data matrices were generated using the R function `mvrnorm`.

Hub correlation structure

The hub correlation structure (referred to as hub observation model) describe the situation where k groups of variables are presented and the observations within each group are correlated with a single observation (the so-called *hub*) with decreasing strength. The k groups are independent, that is there is no correlation among variables belonging to different groups. Set the first observation in each group to be the hub-observation, the correlation $\Sigma_{1,i}$ between variable $i = 1, 2, \ldots, g$ and the hub-observation

$$\Sigma_{1,i} = \rho - \left(\frac{i-2}{g-2}\right)^{\gamma} (\rho - \rho_{min}) \,. \tag{3.32}$$

We simulated a hub correlation structure with 2 groups of unequal size (15 and 5, respectively) and varied ρ between 0.1 and 1.0 in steps of 0.1 using a quadratic attenuation ($\gamma = 2$). Given ρ, random hub-correlation matrices were generated using the R function `simcor.H` provided by Hardin **(Hardin et al. 2013)**. The parameters used were $k = 2$, $\epsilon = 0.01$, $\gamma = 2$, size = (5,2) and edim = 2. Data matrices were generated using the R function `mvrnorm`.

Average

Random correlation matrices Σ_p (with elements ρ_{ij}) were generated satisfying the property

$$\frac{2}{p^2 - p} \sum_{i>j} |\rho_{ij}| = \rho \,, \tag{3.33}$$

that is the average correlation in Σ_p is ρ, having all variables different degree of correlation. This was accomplished by using the vine method **(Ghosh and Henderson 2003; Lewandowski et al. 2009)**. Briefly, correlations are obtained by sampling from a Beta distribution with support $-1 \leq x \leq 1$. The mean μ and the variance σ^2 of the Beta distribution are related to the two Beta shape parameters α and β by the relationships

$$\begin{aligned} \mu &= \frac{\alpha}{\alpha + \beta} \\ \sigma^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \,, \end{aligned} \tag{3.34}$$

from which it follows

$$\begin{aligned} \alpha &= \frac{1}{\sigma^2} - \frac{1}{\mu} \\ \beta &= \alpha\left(\frac{1}{\mu} - 1\right) . \end{aligned} \tag{3.35}$$

The mean μ was numerically optimised to give average correlation ρ between 0.1 and 0.8 in steps of 0.1. The variance σ^2 of the Beta distribution was set to 0.1 in all cases. The corresponding optimised μ values were 0.113, 0.116, 0.123, 0.135, 0.163, 0.201, 0.262, and 0.382, respectively, from which the Beta shape parameters α and β were calculated using Equations (3.35) and used in the generating vine algorithm (see Section 2.4 in **(Lewandowski et al. 2009)**).

3.2.6 Data generation using a dynamic metabolic model

To generate data showing correlation patterns similar to those that can be expected in a standard metabolomic experiment used a dynamic kinetic model. We choose a dynamic model describing the lipopolysaccharide -induced activation of Nuclear Factor kappa B signalling pathway (NF-$\kappa\beta$, Nuclear Factor kappa-light-chain-enhancer of activated B cells). The model consists of 59 ordinary differential equations describing the reactions involving 35 metabolites. The model describes the intra-cellular signalling pathway that activates NF-$\kappa\beta$ p65-p50 in response to lipopolysaccharide which is a gram-negative bacterial endotoxin that triggers an inflammatory response in many cells including uterine smooth muscle cells. The model was obtained from the BioModels database **(Malik-Sheriff et al. 2020)** (www.ebi.ac.uk/biomodels/) with accession number BIOMD0000000489. Full details on the model building and accessory files can be found in the original publication **(Sharp et al. 2013)**.

Simulation of individual metabolite concentration profiles

Subject-specific profiles were generated by varying the Km_i and the k_i constants for all the 59 reactions and the initial concentrations c_m for 4 metabolites with non-zero initial concentrations in the model. The Km_i and the k_i constants and the initial concentrations c_m were sampled from an uniform distribution $\approx U(a, b)$ with lower and upper bounds a and b set to the reference values ±10% as given in the original publication **(Sharp et al. 2013)**.

For j-th individual, the values of k, Km and c for any given reaction were defined as

$$\begin{aligned} k_i^j &\approx U(0.9 \times k_i, 1.1 \times k_i) \,, \\ Km_i^j &\approx U(0.9 \times Km_i, 1.1 \times Km_i) \,, \\ c_m^j &\approx U(0.9 \times c_m, 1.1 \times c_m) \,. \end{aligned} \tag{3.36}$$

We generated 1000 individual profiles from which we randomly sampled data set of varying size ($n = 10, 25, 50, 100, 250$ and 500). In our comparative study we used these data as data set(s) $\mathbf{X}_2$, *i.e.* as reference data set $\mathbf{X}_2$ (see Figure 3.2 for Group (condition) 2. Data for Group (condition) 1 was constructed by varying the values of k_i^j, Km_i^j and c_m^j specific for the j-th individual defined in Equation (3.37) as

$$\begin{aligned} \widetilde{k}_i^j &= \epsilon \times k_i^j \,, \\ \widetilde{Km}_i^j &= \epsilon \times Km_i^j \,, \\ \widetilde{c}_m^j &= \epsilon \times c_m^j \,. \end{aligned} \tag{3.37}$$

where ϵ is a scaling parameter, equal for all subjects and reactions). We varied ϵ over the values $\frac{1}{10}$, $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{1}{1.5}$, 1, 1.5, 2, 3, 5, 10 which were used to generate subject-specific metabolite profiles as described above. Data was collected in data sets $\mathbf{X}_1$ of varying size ($n = 10, 25, 50, 100, 250$ and 500) and for each ϵ value.

3.2.7 Experimental data

We considered the metabolomic data set compendium compiled by Mendez and coworkers **(Mendez et al. 2019)**. The compendium contains 10 data sets representative of the three most common metabolomic experimental platforms (nuclear magnetic resonance NMR; gas chromatography mass spectrometry, GC-MS; liquid chromatography mass spectrometry, LC-MS) applied to metabolomic profiling of different biofluids (urine, serum/plasma, faeces). All the data sets pertain case/control studies with a a clear binary outcome available to model (either a primary or secondary outcome of the publication, or a subset of a multi-class study) and have different sample size and number of variables (metabolites) acquired. Data sets characteristics and references are given in Table 3.1. We made use of the processed cleaned data made accessible via the github link provided in **(Mendez et al. 2019)** and available in xlsx format. We refer to **(Mendez et al. 2019)** for more details about the data processing and cleaning. Data were used as provided by **(Mendez et al. 2019)** with the exception for those data sets where missing data was present: variables with missing data were either removed (data set MTBLS136) or imputed (data set ST001047) using the random forest-based approach implemented in the R package `missForest` **(Stekhoven et al. 2011)**.

In addition we considered other data sets to include also tissues (fat) and plant and fruit extracts together with microbiome data (16S sequencing) and chemical assays on diverse fluids like oil, wine and coffee. For completeness we also included two transcriptomic data sets. Data were derived from the original publications or from R packages with which they were distributed, as indicated in Table 3.1.

The transcriptomic data set were analysed considering only the 250 most and less differential expressed genes between the two classes. Some data sets presented unbalanced groups and they were analysed retaining the original sample size or making them balanced (see Table 3.1 for more details).

3.2.8 Software

Calculations were performed using R **(R Core Team 2013)**, Matlab **(MATLAB 2018)** and Python **(Python Core Team 2015)**. The R code for differential network analysis is available at www.systemsbiology.nl under the software tab.

3.3 Results

3.3.1 Differential connectivity analysis on experimental data

As anticipated in the Introduction we observed a marked difference when calculating the metabolite differential connectivity (see Equations (3.23) and (3.24)) from metabolite-metabolite association network estimated from blood samples collected from male and females subjects (data set no. 15 in Table 3.1) **(Rist et al. 2017)**).

Subsequently, we re-analysed 23 publicly available data sets pertaining to metabolomic studies from different fields, from cancer to plant biology. Although different in scope, most studies followed the same simple experimental design: samples were collected from two groups of subjects or from different conditions with the aim of comparing

Table 3.1: *Correlation* and *MI* indicate the number of features found to be statistically significantly differentially connected (at the $\alpha = 0.05$ level using correlation and Mutual Information as measure of association). *Only in correlation* and *Only in MI* denote differentially connected features found only using correlation and Mutual Information, respectively. *Overlap* indicates those found by both methods. The number of observation (No. observations) is $n = n_1 + n_2$ where n_1 and n_2 is the sample size of group 1 and 2, respectively. Study IDs starting with MTBL indicate data available in Metabolights database **(Haug et al. 2019)** (www.ebi.ac.uk/metabolights) while those starting with ST indicate data available in the Metabolomics Workbench database **(Sud et al. 2015)** (www.metabolomicsworkbench.org). Data set No. 27 has been obtained from the RAST database **(F. Meyer et al. 2008)** (www.mg-rast.org). Data sets without study ID have been derived either from the original publications or from R packages within which they were distributed: BioMark **(Wehrens et al. 2012)**, kodama **(Cacciatore et al. 2017)**, MixOmics **(Rohart et al. 2017)** and pgmm **(Mcnicholas et al. 2008)**. Abbreviations: CD, Chron's disease; CFS, Cronic fatigue syndrome; E Estrogen; E+P, Estrogen + Progesterone; ES, Ewing sarcoma; IBD, Inflammatory bowel disease; MA, microarray; RMS, Rhabdomyosarcoma; UC, Ulcertive colitis. For data set 24 and 25, the superscripts '+' and '-' indicate the 250 most (the least, respectively) expressed genes' and the superscript *r* indicates a random selecion of 500 genes.

								No. deferentially connected features				
No.	Study ID	Ref.	Platform	Type	No. observations	No. features	Design	Correlation	MI	Only in Corr	Only in MI	Overlap
1	MTBLS90	**(Ganna et al. 2014)**	LC–MS	Plasma	968 (485/483)	189	Sex (M/F)	132	101	68	37	64
2	MTBLS92	**(Hilvo et al. 2014)**	LC–MS	Plasma	253 (142/111)	138	Chemotherapy (before/after)	138	12	126	0	12
3	MTBLS136	**(V. L. Stevens et al. 2018)**	LC–MS	Serum	668 (337/331)	371	Homone (E/E+P)	255	125	167	37	88
4	MTBLS161	**(C. W. Armstrong et al. 2015)**	NMR	Serum	59 (34/25)	30	CFS (case/control)	14	12	6	4	8
5	MTBLS404	**(Thévenot et al. 2015)**	LC–MS	Urine	184 (101/83)	120	Sex (M/F)	105	58	51	4	54
6	MTBLS547	**(X. Zheng et al. 2017)**	LC–MS	Caecal	97 (46/51)	35	High fat diet (case/-control)	35	4	31	0	4
7	ST000369	**(Fahrmann et al. 2015)**	GC–MS	Serum	80 (49/31)	181	Adenocarcinoma/He	181	69	112	0	69
8	ST000496	**(Sakanaka et al. 2017)**	GC–MS	Saliva	100 (50/50)	69	Debridement (pre/-post)	59	31	32	4	27
9	ST001000	**(Franzosa, Sirota-Madi, et al. 2019)**	LC–MS	Stool	121 (68/53)	124	IBD (CD/UC)	96	79	33	16	63
10	ST001047	**(A. W. Chan et al. 2016)**	NMR	Urine	83 (43/40)	149	Gastric cancer/healthy	109	85	42	18	67
11	ST000061		GC-MS	Tissue	118 (59/59)	157	subcutaeus/visceral fat	156	83	73	0	83
12		**(Eisner et al. 2011)**	NMR	Urine	50 (25/25)	200	cachexia (case/control)	163	57	115	9	48
13		**(Eisner et al. 2011)**	NMR	Urine	77 (47/30)	63	cachexia (case/control)	63	33	30	0	33
14		**(Eisner et al. 2011)**	NMR	Urine	60 (30/30)	63	cachexia (case/control)	55	43	15	3	40
15		**(Rist et al. 2017)**	GC-MS	Plasma	291(172/119)	128	Sex (M/F)	128	23	105	0	23
16		**(Rist et al. 2017)**	GC-MS	Plasma	200 (100/100)	128	Sex (M/F)	103	51	56	4	47
17		**(Rist et al. 2017)**	GC-MS	Urine	301 (129/172)	324	Sex (M/F)	256	143	136	23	120
18	MTBLS123	**(Lusczek et al. 2013)**	NMR	Urine	151 (79/72)	63	Shock (pre/post)	63	9	54	0	9
19	ST001243	**(Powers et al. 2019)**	GC-MS	Plasma	98 (48/50)	69	Trisomy 21 (yes/no)	69	28	41	0	28
20	MTBLS147	**(Vignoli, Tenori, Luchinat, et al. 2017)**	NMR	Plasma	370 (185/185)	417	Sex (M/F)	417	414	3	0	414
21	KODAMA	**(Bernini et al. 2009)**	NMR	Urine	80(40/40)	490	Subject (A/B)	459	293	187	21	272
22		**(Caldana et al. 2011)**	GC-MS	Plant	70 (35/35)	67	Light/Dark	37	19	22	4	15
23	BioMark	**(Wehrens et al. 2012)**	LC–MS	Apple	20 (10/10)	198	Treated/Untreated	124	58	83	17	41
24	MixOmics	**(Khan et al. 2001)**	MA	Cell	43 (23/20)$^-$	250	Sarcoma (RMS/ES)	250	18	232	0	18
25	MixOmics	**(Khan et al. 2001)**	MA	Cell	43 (23/20)$^+$	250	Sarcoma (RMS/ES)	250	8	242	0	8
26	MixOmics	**(Bushel et al. 2007)**	MA	Cell	32 (16/16)^r	500	High/Low dose	405	279	170	44	235
27	4537568.3-776.3	**(Stanley et al. 2013)**	16S seq	Faeces	145 (71/74)	243	Flock (A/B)	241	150	91	0	150
28	pgmm	**(forina1983classification)**	Chemical assay	Oil	50 (25/25)	7	Region (A/B)	4	0	4	0	0
29	pgmm	**(Streuli 1973)**	Chemical assay	Coffee	43 (36/7)	12	Variety (Arabica/Robusta)	4	11	0	7	4
30	pgmm	**(Forina, Armanino, Castino, et al. 1986)**	Chemical assay	Wine	130 (59/71)	27	Type (Barolo/Grignolino)	8	10	5	7	3

Table 3.2: Results of pathway enrichment for data set 12 and 25 from Table 3.1 based on the sets of metabolite found to be differentially connected using correlation or Mutual information as measure of metabolite-metabolite association. FDR: False discovery rate. Empty cells indicate that the no metabolite was found to be associated with the given pathway.

	Pathway enrichment based on			
Data set 12	**Correlation**		**Mutual Information**	
Pathway	Raw P	FDR	Raw p	FDR
Aminoacyl-tRNA biosynthesis	3×10^{-12}	3×10^{-12}	0.0006	0.05
Valine, leucine and isoleucine biosynthesis	3×10^{-5}	0.001		
Alanine, aspartate and glutamate metabolism	6×10^{-5}	0.002		
Arginine biosynthesis	0.0004	0.008	0.006	0.18
Glyoxylate and dicarboxylate metabolism	0.001	0.020	0.25	1.00
Glycine, serine and threonine metabolism	0.002	0.020	0.03	0.72
Citrate cycle (TCA cycle)	0.002	0.020		
Phenylalanine metabolism	0.002	0.020	0.09	0.91
Phenylalanine, tyrosine and tryptophan biosynthesis	0.004	0.040		
	Pathway enrichment based on			
Data set 25	**Correlation**		**Mutual information**	
Pathway	Raw P	FDR	Raw p	FDR
Citrate cycle (TCA cycle)	5×10^{-5}	0.004		
Alanine, aspartate and glutamate metabolism	0.0004	0.016	0.15	1
Glyoxylate and dicarboxylate metabolism	0.001	0.020	0.17	1
Glycine, serine and threonine metabolism	0.001	0.020	0.18	1
Histidine metabolism	0.002	0.036	0.09	1
Tyrosine metabolism	0.004	0.050		

profiles between group 1 and group 2. A list of the data sets considered is given in Table 3.1 together with a summary of sample size, number of metabolites measured, the experimental platform and the study design.

For each study we calculated a weighted adjacency matrix using both Pearson's correlation and Mutual Information via empirical estimation for the two groups and for each metabolite we defined the weighted connectivity which was compared between the two groups defined by the study design and whose significance was assessed using a permutation test as illustrated in Figure 3.2. Results are shown in Table 3.1. In all cases the number of differentially connected metabolites (at an $\alpha = 0.05$ confidence level) was much higher when correlation was used as a measure for association and subsequently used to calculate the metabolite connectivity.

This has of course tremendous implications for data interpretation. For instance, if differentially connected metabolites are used for enrichment and/or pathway analysis, a great deal of information may be lost. Consider for instance data set 12 in Table 3.1, which collects GC-MS metabolite profiles of healthy men and women. If pathway analysis is performed on the differentially connected metabolites found using correlation or Mutual Information, the results are strikingly different: only one pathway (Aminoacyl-tRNA biosynthesis) is found to be enriched (FDR < 0.05) when using mutual information as a measure of association. Eight pathways are found only when using correlation. Results are shown in Table 3.2. A similar exercise can be performed for data set no. 25 in Table 3.1. In this case, there is no pathway enriched when using Mutual Information.

On the basis of this analysis we could not draw unequivocal conclusions. In general there is overlap between the metabolites found to be differentially connected using correlation or Mutual Information, but in many cases metabolites are found to be differentially connected only when using one of the two measures. For instance, for data set 1 in Table 3.1, we observed 132 metabolites out of 189 to be differentially connected when using correlation and 90 when using MI, with 64 found with both measures; however 68 metabolites were found only with correlation and 37 only with MI.

To investigate if these patterns were specific to metabolomic data, we analysed with the same approach three transcriptomic data sets, one microbiomics data set and three data sets pertaining to chemical assays. With the exception of data set 29 and 30 we again observed more differentially connected metabolites when using correlation.

Most data sets are unbalanced, with one group larger than the other: we reanalysed some of the data sets by making them balanced to remove this possible confounding factor. This did not affect the results, which were qualitatively the same: the use of correlation resulted in more differentially connected metabolites also when data is balanced.

3.3.2 Type I error

Given the results on experimental data, we questioned our validation procedure based on permutation, speculating that the permutation test based on correlation could have resulted, for some reason, in an inflated Type I error, leading to false positives.

To assess this we devised a simulation strategy where groups 1 and 2 (see Figure 3.2) were substituted with uncorrelated random data generated under a multivariate normal model, which implies that no variable (metabolite) is differentially connected. Under this simulation scheme, the observed number of differentially connected metabolites should be around 5 (*i.e.* 5% of the total number of metabolites tested, if the significance test is performed at $\alpha = 0.05$ level).

We recorded the Type 1 error as a function of sample size n, varying n from 25 to 500. As shown in Figure 3.4, the observed Type I error is always around 0.05, independent from the sample size, and from the particular measure of association used. On the basis of this, we could exclude the possibility of inflated Type I error when correlation was used.

3.3.3 Comparison of correlation and mutual information on simulated data with known correlation structure

We set up a strategy to investigate the behaviour of correlation and Mutual Information for differential network analysis further. We generated data with known correlation structures as detailed in Sections *Toeplitz correlation structure*, *Hub correlation structure* and *Average* and confronted them with data with uncorrelated structures. The number of variables (*i.e.* metabolites) was fixed to 20 while the number of samples varied between 10 and 500. In all cases we varied the strength of the correlation ρ between 0 and 1 which means that apart from the case $\rho = 0$.

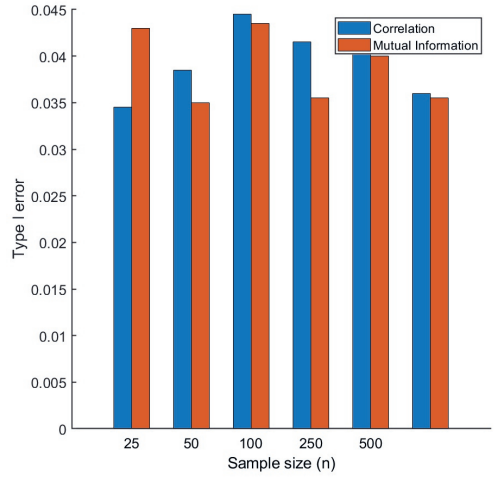


Figure 3.4: Type I error for the permutation test used to assess the statistical significance of metabolite connectivity. Two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ are generate of size $n \times 20$ under an uncorrelated multivariate model ($\mathbf{X}_1 \approx N(0, \mathbf{I})$. Differential connectivity is calculated as described in Equations (3.23) and (3.24) and assessed with a permutation test at the $\alpha = 0.05$ significance level. The overall procedure is repeated 100 times.

In this case we used the four entropy estimators outlined in Sections 3.2.3, 3.2.3, 3.2.3 and 3.2.3 to investigate if the particular choice of a method to estimate the entropy necessary to calculate the Mutual Information had any effect on the estimation of differential connectivity. Overall we did not observe any relevant difference when using different methods and for this reason we present and discuss only the results obtained using the empirical probability distribution to estimate the entropy (see Equation (3.14)). Results are shown in Figure 3.5.

In all cases, we found more differentially connected metabolites using correlation indices as a measure for association than any of the four Mutual Information methods. As it is to be expected, the number of differentially connected metabolites varied with both the sample size and the magnitude of the known correlation ρ of the correlation structures. It should be noted that in our simulation scheme, the differential connectivity is always tested under the alternative hypothesis (see Equation (3.29)) being true (except when $\rho = 0$) and thus the significant differential connectivity in every situation is expected to be 20 for $\rho = 0.1$ to 1.0 and 0 for $\rho = 0$.

The general trend seen in analysing the number of significantly differentially connected metabolites increases with both sample size n and the known correlation ρ of the data structures. As for any statistical test, the power of our approach increases with both sample size and effect magnitude. We notice that at $n = 500$ and $\rho > 0.8$, most methods display the significance of differential connectivity to be 20 with any of the data structures we tested against.

Mutual information is only able to show significant differential connectivity of 20 at $\rho > 0.8$ irrespective of the sample size, indicating a reduction of power to detect differential connectivity. Interestingly, we observed that the performance of Mutual Information in inferring the differential connectivity drops significantly at $\rho = 0.3$

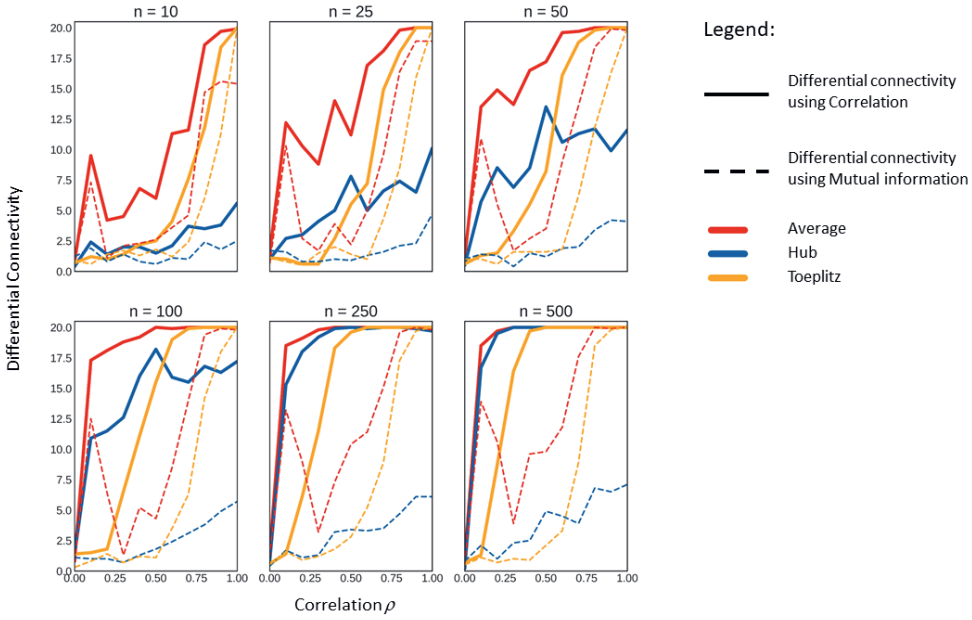


Figure 3.5: Median of the significant differentially connected variables on all simulated data sets per known correlation ρ per sample size n

and then trends upwards again. This observation was consistent for all sample sizes and all methods used to estimate the entropy in this study.

In all cases we observed the maximal differential connectivity (*i.e.* 20) is always achieved for smaller values of ρ and smaller sample size when using correlation rather than Mutual Information.

Given the above mentioned hypothesis, it might be easier to understand why when Mutual Information is used as the measure for association, it performs extremely poorly in identifying differential connectivity. The poor performance is not affected by sample size or by the underlying data correlation structure. These results confirm what was observed when analysing real life metabolomics data set.

3.3.4 Comparison of correlation and mutual information on simulated data from a dynamic model

The dynamic metabolic model of the NF-$\kappa\beta$ was used to generate physiologically plausible metabolite concentration profiles for n individuals as detailed in **(Jahagirdar, Suarez-Diez, et al. 2019)**, mimicking the real-life process of data generation from a population of subjects. This data presents metabolites with complex, non-linear relationships that are almost impossible to simulate with statistical methods; hence this approach gives a better representation of the metabolite-metabolite association patterns observed in real-life experimental data.

Working in a two-groups scenario (see Figure 3.1 and 3.2) we varied the kinetic parameters using the multipliers (ϵ) to change the behaviour of the entire model. The

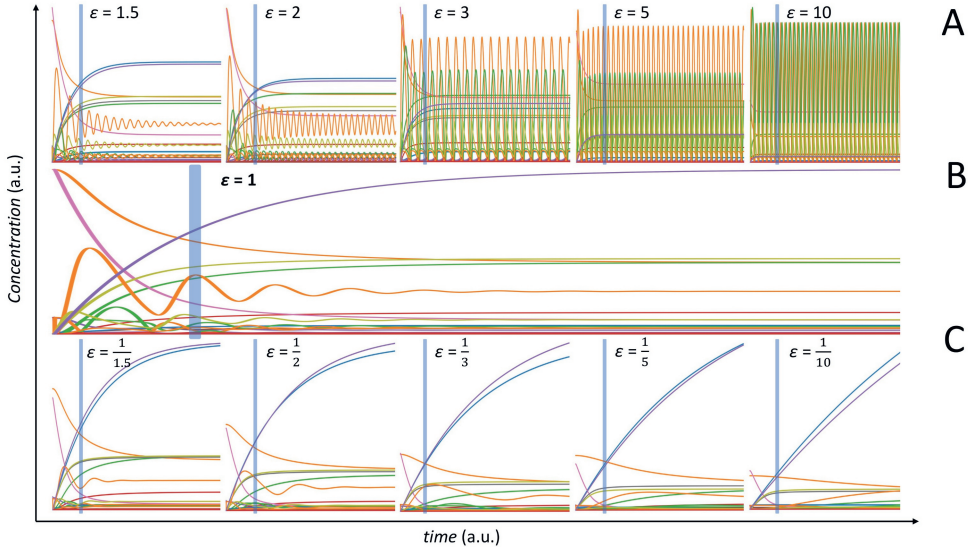


Figure 3.6: Behaviour of the NF-$\kappa\beta$ dynamic model. A) Time concentration profiles for model perturbation with $\epsilon > 1$. B) Original model. C) A) Time concentration profiles for model perturbation with $\epsilon < 1$. Different colours correspond to different metabolite time profiles. The vertical lines indicate the time sampling point.

effect of the modification of the kinetic parameters on the overall model behaviour is shown in Figure 3.6. Values of $\epsilon > 1$ induces fast oscillations in the concentration profiles of certain metabolites (panel A), while values of $\epsilon < 1$ flattens out the oscillating behaviour (panel C). Panel B of Figure 3.6 gives the time concentration profiles for the original, unperturbed, model.

Here we used ϵ as a measure of the perturbation of the dynamic model (data in $\mathbf{X}_1$) with respect to the original one defined under normal physiological conditions (data in $\mathbf{X}_2$). However it should be noted that it is difficult to relate ϵ to the number of possibly differentially connected metabolites. This is because it is not possible to predict the relationship among metabolites directly from the structure of the dynamic model. As a matter of fact, the use of the dynamic metabolic model allows a more exhaustive analysis on metabolite associations but correlations observed in the data do not always reflect the structure of the metabolic network: two metabolites can be direct neighbours in the metabolic network but not correlated; conversely two metabolites can be very distant in the metabolic network but showing high correlation.

The connectivity is formally tested under a null hypothesis scenario like in the case of data generated under different correlation models (see Sections *Toeplitz correlation structure*, *Hub correlation structure* and *Average*) but in this case, the expected connectivity for each metabolite in the NF-$\kappa\beta$ model for the unperturbed case ($\epsilon = 1$) is different from 0.

Also in this case, the use of correlations results, on average, on more differentially connected metabolites, than when using Mutual Information, as shown in Figure 3.7. Pearson's and Spearman's correlation performed similarly for most cases and

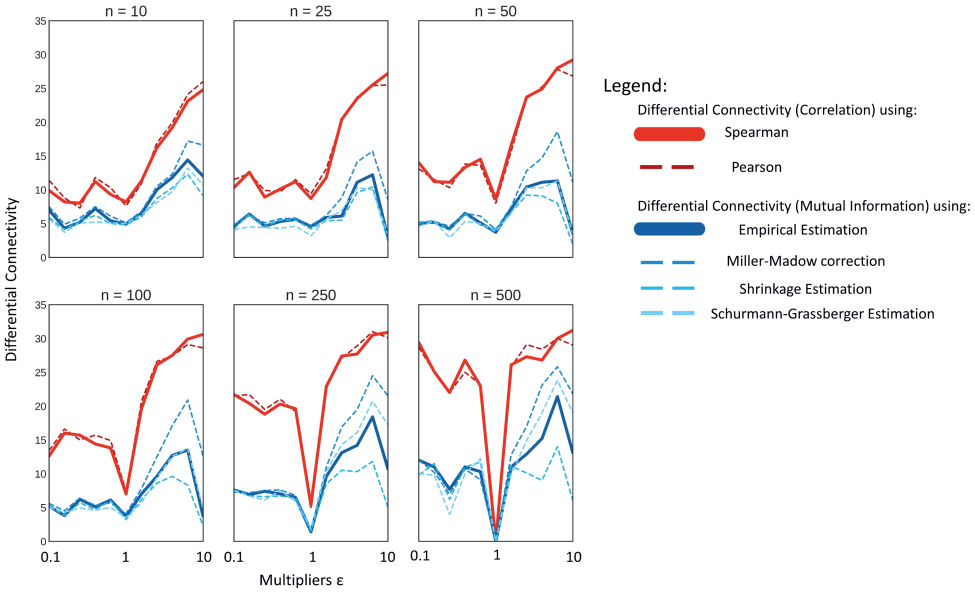


Figure 3.7: Median of the significant differentially connected variables on data simulated using the NF-$\kappa\beta$ dynamic model as a function of the model perturbationϵ and the sample size n.

the marginal difference of Pearson correlation performing better in extremely low sample sizes might be explained by the bias created between the relationship of the two correlation methods as shown in Figure 3.10 in Section *Relationship between correlation and mutual information*.

There is an inherent difference in the change of behaviour in the model with $\epsilon < 1$ and $\epsilon > 1$ as shown in Figure 3.6. There is a significant increase in oscillations, at least for some metabolites, when $\epsilon > 1$ with the magnitude and the frequency of the oscillations increasing with ϵ. This introduces high non linearity in the data and may partially explain why Mutual Information performs better with $\epsilon > 1$ than with $\epsilon < 1$. However this does not explain the differences observed between correlation and Mutual Information.

We observed the differential connectivity to be zero for $\epsilon = 1$ only for large sample size $n = 500$, suggesting the existence of spurious associations for small sample size and/or instability in the estimation of both correlation and Mutual Information.

We speculate that the perturbation in the kinetic parameters may induce pseudo-associations among metabolites that are picked-up by correlation but not by Mutual Information, thus increasing metabolites connectivity (see definition in Equation (3.24)). These pseudo-associations may be stronger when $\epsilon > 1$ and the system is oscillating with high frequency, since small changes in kinetics can result in larger variation in concentration when sampling happens at a constant time as in the present case. When $\epsilon < 1$ most metabolites exhibit smooth linear and exponential curves and the variability in concentration is greatly reduced. For example, consider two metabolites M1 and M2, with the concentration of M1 following an exponential curve for $\epsilon = 1$ and $\epsilon > 1$, while M2 shows a small oscillation behaviour with $\epsilon = 1$ and a

large oscillation with $\epsilon > 1$. If sampling happens at, say, $t = 10000$ units, at $\epsilon = 1$ there would be small variations in M1 and M2; however at $\epsilon > 1$ there might be large variations in M2 depending on whether the crest or the trough is picked up, especially if the frequency and amplitude are high. This would result in a situation where when $\epsilon = 1$ small change in M1 is correlated to small change in M2 and when $\epsilon > 1$ small change in M1 is correlated to a large change in M2 and hence the two variable would show up as differentially connected when the relationship change between them might be less subtle. As the number of samples is increased, the occurrence of such pseudo-associations will be reduced.

In contrast with what was observed with data generated under different correlation models, we observed differences when using different algorithms for the estimation of Mutual Information. In particular the asymptotic bias was large and observable. Indeed, using the Miller-Madow correction (see Section *Miller-Madow asymptotic bias corrected empirical estimator*) resulted in a marked increase in the performance of Mutual Information especially with $\epsilon > 1$. On the contrary, the shrinkage estimation of entropy failed to show any increase in performance for inferring differential connectivity as the sample size was increased, confirming previous observations that the shrinkage estimation is more effective at lower sample sizes **(Nemenman et al. 2004)**.

When using correlation, for small sample size ($n \leq 50$) the number of differentially connected metabolites for the case of data generated with $\epsilon < 1$ seems not to vary while it increases for $\epsilon > 1$. For larger sample size ($n \geq 250$) the number of differentially connected metabolites exhibits a symmetric behaviour with respect to $\epsilon = 1$. A similar behaviour is observed when using Mutual Information, which shows less sensitivity to detect differentially connected metabolites, especially for $\epsilon < 1$ and small sample size. The sub-optimal performance of Mutual Information to infer connectivity can be explained by considering the analytical relationship existing between Pearson correlation and Mutual Information, as shown in Section *Relationship between correlation and mutual information*.

3.3.5 Relationship between correlation and mutual information

In the case of two bivariate variables x_1, x_2 linearly correlated with correlation ρ, there is a direct relationship between the Mutual Information $MI(x_1, x_2)$ and ρ. If

$$(X_1, X_2) \approx N(\boldsymbol{\mu}, \Sigma) \,, \tag{3.38}$$

with

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_{12} \\ \rho\sigma_{12} & \sigma_2^2 \end{pmatrix}, \tag{3.39}$$

where σ_1^2 and σ_2^2 is the variance of x_1 and x_2, respectively and σ_{12} their covariance, it holds that (see Equation 2.8 in **(Gelfand et al. 1957)**):

$$MI(X_1, X_2) = -\frac{1}{2}\log(1 - \rho^2) \,. \tag{3.40}$$

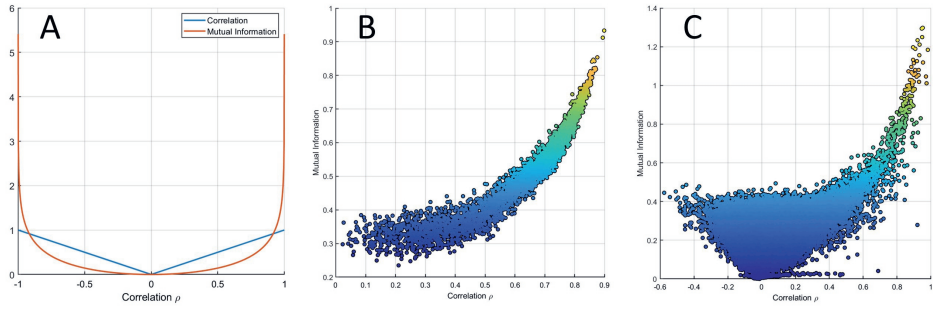


Figure 3.8: A) Mutual Information $MI(X_1, X_2)$ of two bivariate variables X_1, X_2 linearly correlated with correlation ρ as a function of ρ. The two curves intersect at approximately $\rho = 0.916$. B) MI versus Pearson's correlation from data simulated with an average correlation of 0.6 (beta simulation). C) MI versus Pearson's correlation from experimental data (Data set 3 from Table 3.1).

From Equation (3.40) it follows that if two variables are linearly (cor)related, their Mutual Information is (almost) always smaller than their correlation. This is shown in Figure 3.8 where the relationship (3.40) is given for $-1 \leq \rho \leq$: $MI(X_1, X_2)$. In particular, it holds that

$$MI(X_1, X_2) \rightarrow \begin{cases} < \rho & \text{if } |\rho| < 0.916 \\ = \rho & \text{if } |\rho| = 0.916 \\ > \rho & \text{if } |\rho| > 0.916 \end{cases} . \tag{3.41}$$

The relationship between Mutual Information and correlation is shown for data simulated under the average model (see Equation (3.33)) in Figure 3.8B and for experimental data set 3 from Table 3.1 in 3.8C, which show good agreement between the analytical relationship between correlation and Mutual Information given in Equation (3.33). Figure 3.9 shows the same relationship for data generated using the NF-$\kappa\beta$ dynamic model.

A similar behaviour is also observed when Spearman's correlation is used as an index of association. In fact, if there are no ties, the Pearson's and Spearman's correlation coefficient are related, for sample size n, by the formula **(Kendall 1948)**

$$\rho_S = \frac{6}{\pi(n+1)} \left[\arcsin \rho + (n-2) \arcsin\left(\frac{\rho}{2}\right) \right] , \tag{3.42}$$

which is shown in Figure 3.10. For linearly positively correlated variables as in the present simulation, Spearman's correlation is biased downwards (in absolute value)

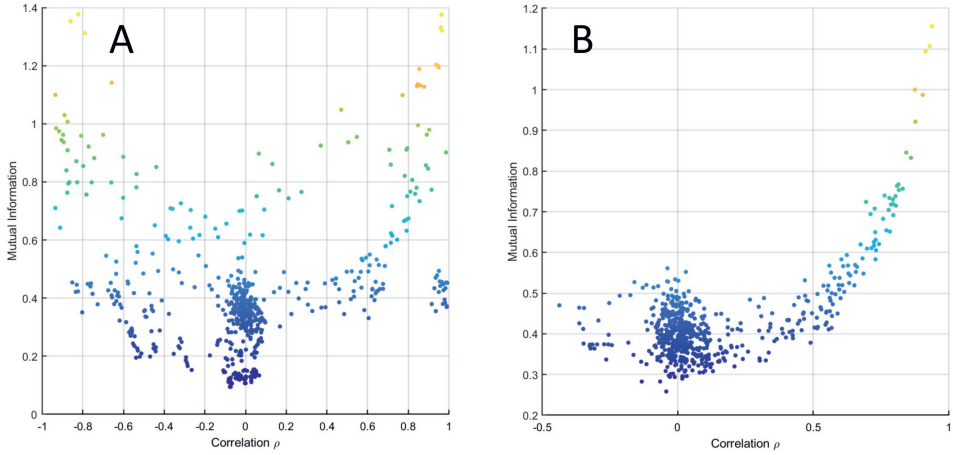


Figure 3.9: MI versus Pearson's correlation from data simulated with the NF-$\kappa\beta$ dynamic model when with A) $\epsilon = 0.1$ and B) $\epsilon = 10$.

and the difference is maximal for $\rho = 0.577$ (respectively, for $\rho = -0.577$ for negatively correlated variables.). The magnitude of the bias depends on the sample size n, but the location where it assumes maximum value is independent from n. For a calculation see reference **(Zimmerman et al. 2003)**. However for large sample size ($n > 50$) the bias introduced by taking the Spearman's correlation in place of the Pearson's to quantify association is negligible and as a consequence the estimation of the differential connectivity is not affected.

3.4 Discussion

Correlation and mutual-Information measures have been widely used in many research applications to quantify and describe the relationships between variables and thus have become the foundations for network inference methods **(Jahagirdar, Suarez-Diez, et al. 2019)**. In general researchers trained in statistics tend to use correlation based indices while researchers trained in computer science gravitate towards mutual-information. However, the use of the correlation coefficient is much more widespread in life sciences research than mutual information: a Pubmed search (March 2020) returned 61709 hits for "correlation coefficient" against and 3582 hits for "Mutual information". Inference methods based on correlation can only detect linearly direct associations and can miss nonlinear relations, which play essential roles in many nonlinear systems, such as biological systems **(J. Zhao et al. 2016)** In this light, Mutual information has attractive properties especially when dealing with the detection of non-linear relationships **(Cover et al. 2012)**. This was one of the main reasons we expected mutual information to have superior performance in metabolite-metabolite association networks given the non-linear nature of the relationships existing among metabolites concentrations. Being based on mutual independence, mutual information can be considered to be a nonlinear version of

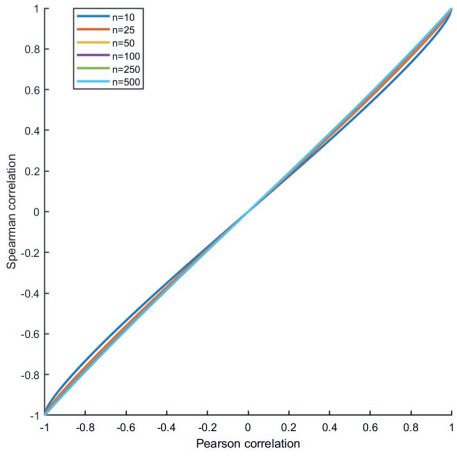


Figure 3.10: Relationship (see Equation (3.42)) between the Spearman's (3.2) and the Pearson's (3.1) correlation coefficients for linearly correlated data for different sample size n.

correlation that can detect nonlinear correlations (but not direct associations or dependencies owing to the information of only joint probability), and have the same overestimation problem as correlation **(J. Zhao et al. 2016)**.

Correlation and mutual information measure have been compared mostly in the framework of gene networks inferences. Steuer *et al.* showed an almost one-to-one correspondence between correlation and mutual information when measuring gene pairwise relationships **(Steuer et al. 2002)**, while Lindolf *et al.* found no superior merits of mutual information for constructing co-expression networks **(Lindlöf et al. 2005)**. Song *et al.* examined different correlation-based measure of association and found them to outperform MI in terms of elucidating gene pairwise relationships **(Song et al. 2012)**. In gene ontology studies it has been observed that, when robust correlation and robust mutual-information has disagreed, the robust correlation findings seemed to be statistically and biologically more plausible **(Song et al. 2012)**.

There is little literature on the use of mutual information in metabolomics applications (12 hits for a Pubmed query "metabolomics AND mutual information", performed in March 2020). Numata *et al.* found that mutual information was able to detect additional non-linear correlations undetectable for the Pearson coefficient **(Numata et al. 2008)** and Yu *et al.* concluded that Spearman and MI indexes outperform the other measures to co-associate metabolite and microbiome data **(You et al. 2019)**. Based on **(Kraskov et al. 2004; Matsuda 2000)**, Numata *et al.* also advocated for the use of mutual information since mutual information for pairs of variables is not altered by homeomorphic (non-linear) transformations of the data, which may be relevant because metabolomic data rarely yield absolute concentrations, but rather ratios of concentrations **(D. Camacho et al. 2005)**. However, Saccenti *et al.* found mutual information to overestimate chance associations **(Saccenti, Suarez-Diez, et al. 2015)**. Correlation are objectively difficult to estimate and are sensitive to experimental noise **(Edoardo et al. 2020)** and to data pre-processing like normalization **(Saccenti 2017)**. However correlation indexes have nice properties such as *i*) it can

be easily calculated, *ii*) it allows for asymptotic statistical tests (regression models, Fisher transformation) for calculating significance, and *iii*) the sign of correlation allows one to distinguish between positive and negative relationships.

Although in this study we ignored the directionality of the relationships to build networks and calculate connectivity and perform connectivity analysis, this is a an inherent limitation of mutual information that cannot capture directionality and changes thereof since it is a strictly semi-positive quantity **(Mason et al. 2009)**. In fact, (strong) positive correlation can indicate an equilibrium condition or enzyme dominance, while strong negative correlation can indicate the presence of a conserved moiety **(D. Camacho et al. 2005)**. In addition, correlation indices can be calculated with significantly fewer samples than mutual information **(Song et al. 2012)** and we observed Mutual information to require significantly larger sample sizes to obtain the same robustness attained by correlation. Moreover, the estimation of mutual information depends on the particular choice of algorithms and user defined parameter setting **(Doquire et al. 2012)** and we also observed dependence on the estimation algorithm when mutual information is used for differential connectivity analysis.

On the basis of our investigation concerning the use of correlation and mutual information for differential connectivity analysis we can conclude that *i*) Pearson's and Spearman's correlation coefficient are better to detect deferentially connected metabolites than mutual information methods in metabolite-metabolite association networks created from experimental data, simulated data with known correlated structures and from a dynamic metabolic model; *ii*) When a dynamic metabolic model was used to simulate real-world like observational data, different methods to estimate entropy showed different performance. However the same could not be concluded when simulated data structures were used. *iii*) When analysing the relationship between correlation and mutual-information, we find that mutual-information of two linearly related variables is almost always less than that of their correlation and this was observed in real metabolomics data, simulated data and data simulated using a the NF-$\kappa\beta$ dynamic model.

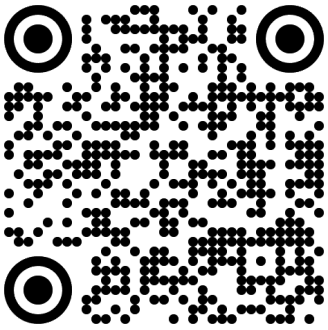
Overall, the present investigation indicates that there is no benefit in using mutual information in place of standard Pearson's and Spearman's correlation when the focus of the application is the detection of differentially connected metabolites in differential network analysis.

3.5 Author contributions

Conceptualization, E.S.; methodology, S.J., E.S.; software, S.J.; validation, S.J. and E.S.; formal analysis, S.J.; investigation, S.J. and E.S.; resources, E.S.; data curation, E.S.; writing-original draft preparation, S.J. and E.S.; writing-review and editing, S.J. and E.S.; supervision, E.S.; funding acquisition, E.S. All authors have read and agreed to the published version of the manuscript.

3.6 Funding

This study has received funding from The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (number 456008002)



Publication

under the PerMed Joint Transnational call JTC 2018.

Chapter 4

Chapter

Sanjeevan Jahagirdar[1], Yasmijn Balder[1], Oddvar Oppegaard[2], Knut Anders Mosevoll[2,3], Hakon Reikvam[2,3], Trond Bruun[2,3], Mattias Svenson[4], Ole Hyldegaard[5,6], Anna Norrby-Teglund[4], INFECT Study group[#], Vitor A. P. Martins dos Santos[7,8], Steinar Skrede[2,3], Edoardo Saccenti[1]
#INFECT study group (Trond Bruun, Eivind Rath, Torbjørn Nedrebø, Per Arnell, Anders Rosen, Morten Hedetoft, Martin B. Madsen, Mattias Svensson, Johanna Snäll, Ylva Karlsson, & Michael Nekludov)

Turn to page 377 for author affiliations

This chapter is prepared for publication

Eventually it Boils Down to the Clinical Variables

Abstract

Introduction: Necrotising soft tissue infections (NSTI) are severe bacterial infections frequently associated with tissue necrosis, sepsis, and high mortality rates. They are often complicated by rapid progression and vague symptoms. The focus of this study is to comprehend the association structures in order to develop conceptual frameworks that consolidate extensive datasets for a better understanding of the disease and its mechanisms.
Methods: In this investigation, we employ two mixed graphical models (pairwise Markov random field & GRaFo), specifically designed to examine conditional dependencies between mixed data types in the context of limited sample sizes. This approach allows us to analyse the association structures that emerge between plasma analyte concentrations, human and bacterial gene expressions, and variables measured in the clinic. Moreover, we examine the resulting network by identifying network hubs and motifs, highlighting the central variables and their associations.
Results: The central conditional dependencies manifested in the network underline the intricate interactions taking place between *S. pyogenes* and neutrophils. Levels of Thrombomodulin are demonstrated to be associated with markers of severe sepsis and elements such as bacterial virulence factors and toxins, known to interact with the human immune system and blood constituents. Correspondingly, mortality is found to be associated with the regulation of virulence factors and toxins, namely SpeB, Streptolysin S, and Sic, orchestrated by CcpA to oversee biofilm expression.
Conclusion: Employing a systems biology approach, we investigated the association structures of all the measured variables in NSTI patients, uncovering key pathogenic mechanisms involving *S. pyogenes*, neutrophils, and Thrombomodulin, and proposed data-driven hypotheses for future study. These findings have the potential to contribute to the further development of intervention and treatment strategies.

4.1 Introduction

Severe bacterial infections can be a function of several factors that include environmental risk factors, predisposition to the infection, and the intricate interactions that play out between bacterial species and the human immune system **(Doron et al. 2008)**. Necrotising soft tissue infections (NSTI) are one such devastating bacterial infection. They are often characterised by their rapid progression and high risk of mortality coupled with vague symptoms **(Hua et al. 2022; D. L. Stevens and Bryant 2017)**. These rare infections are aggressive in nature causing impairment and injury to soft tissue compartment often resulting in severe losses in the quality of life. **(Urbina et al. 2021; Suijker et al. 2020; M. B. Madsen, Bergsten, et al. 2020)**.

The debilitating symptoms observed in NSTI patients have been shown to be a result of both microbial action and host response **(Siemens, Snäll, et al. 2020; Medina et al. 2021)**. When comprehending intricate infections characterised by overlapping and ambiguous symptoms, the hypotheses frequently extend beyond a solitary host response triggered by a predetermined set of predictors. In the case of complex diseases, the focus sometimes shifts towards comprehending the association structures encompassing multiple omics layers, aiming to develop conceptual frameworks or consolidate extensive datasets for a better understanding of the disease **(Hasin et al. 2017)**. Extracting associations from such data often requires stringent measures that can differentiate between direct and indirect associations. Partial correlations and other similar measures have been traditionally used to estimate these types of conditional dependencies/independencies **(De La Fuente et al. 2004)**.

NSTIs exhibit a relatively low incidence rate (0.2 to 15.5 per 100,000 people/year) **(D. L. Stevens and Bryant 2017)**, which contributes to challenges in our understanding of NSTI due to limited sample sizes. In this study, we leverage the multi-omics data obtained from a prospective multicenter INFECT study conducted on a NSTI patient cohort **(M. Madsen et al. 2018)**. The small sample sizes result in a substantial disparity between the number of measurements per sample and the sample size itself. Consequently, traditional measures like partial correlation become impractical. Additionally, the presence of mixed data types in the measurements further complicates the assessment of such associations.

This study builds upon research investigating dual RNA-seq data **(Thänert et al. 2019; Jahagirdar, L. Morris, et al. 2022)** in the chapter *Inside the Inferno: An Inspection of the Host-Pathogen Interactions* and plasma analytes **(Medina et al. 2021; Rath et al. 2023)** in the chapters *The Race Against Time: Discriminatory Plasma Biomarkers* and *The Immune System Responds: Systemic Immune Activation Profiles*, though neither of these studies established a link between the omics data and parameters measured in the clinic. Rapidly progressive necrotising soft tissue infections pose a significant diagnostic challenge, given the inadequacy of current diagnostic tools **(T. Chan et al. 2008)**. Consequently, it is crucial to establish associations between omics data and readily accessible clinical measurements. Such associations can enable timely diagnosis and treatment while expanding our mechanistic understanding of the underlying processes.

This study is an exploratory discourse where we employ methods capable of handling both discrete and continuous variables to estimate conditional dependence/in-

dependence between clinical variables, plasma analytes, and human and bacterial genes. We further analyse the topologies of the ensuing networks in the form of network hubs and network motifs and evaluate the associations with variables of interest to form hypotheses that can be pursued further.

4.2 Methods and Materials

4.2.1 Patient cohorts

All the data comes through samples collected from surgically confirmed necrotising soft tissue infection patients that were enrolled in the multicenter INFECT study. These samples were from 5 hospitals namely, Blekingesjukhuset (Karlskrona, Sweden), Haukeland University Hospital, Karolinska University Hospital, Righospitalet (Copenhagen, Denmark), and Sahlgrenska University Hospital. In this study, we have categorised the data collected from patients with NSTI into four distinct groups, namely clinical variables, human genes, bacterial genes, and cytokines.

4.2.2 Patient data

Clinical variables consist of measurements that are taken in the hospital. These measurements include patient data, risk factors, affected body parts, admission data, clinical data, variables registered daily in the ICU, medications, blood samples, variables prior to diagnosis, variables regarding treatments, follow-up results and more. The study design behind the clinical variables can be found in **(M. Madsen et al. 2018)** and a complete detailed list of all the clinical variables measured can be found in appendix S1 of the manuscript **(M. Madsen et al. 2018)**. Human and bacterial gene expressions were obtained from dual RNA sequencing of NSTI patient biopsies. Detailed information on the microbial community profiling using 16S rRNA and the bioinformatic analysis of the data can be found in **(Jahagirdar, L. Morris, et al. 2022; Thänert et al. 2019)** and their supplementaries. Cytokine data consists of analyte concentrations measured from blood plasma of the NSTI patients using a customised Luminex multiplex assay. Two analytes(IL-23 & IL-33) were measured using ELISA. The measured analytes include chemokines, interleukins, adhesion molecules, matrix metalloproteases, and other biomolecules. For the purpose of this manuscript, we have sometimes collectively referred to this data as cytokine data, although not all the measured analytes are cytokines. Detailed information on the measurement and a complete list of the analytes can be found in **Chapter 6** and **Chapter 7 (Medina et al. 2021; Rath et al. 2023)**.

Not all patients had data available in all four categories. Plasma analytes were measured on day 0 (day of admission to the hospital) and day 3 (72 hrs post admission) for some patients. As the subset of patients with day 3 was small, we decided to utilise the day 0 measurements only. In this study, we include patients that have data from all four categories (n = 44).

4.2.3 Data imputation

In the clinical variables, we remove all variables that have more than 20% missing values. In the remaining variables, we use missforest, a random forest based imputation method to impute missing data. We use the standard functions from the missforest package **(Stekhoven et al. 2012)** with 1000 trees and 100 iterations.

The cytokine concentrations had missing values mainly due to the values being outside of the measurement equipment range. A detailed imputation strategy was applied to the missing values of cytokine concentrations based on the information on account of the left-centerdness or right-centerdness of the missing values. The detailed strategy is described elsewhere **(Medina et al. 2021)**.

4.2.4 Data transformation

The clinical variables measured in the form of discrete variables are transformed into dummy variables using the One-Hot encoding method.

Due to the difference in the presence of bacteria (and therefore bacterial genes) based on different types of NSTI, there is a need to differentiate between 0 values in the gene transcriptomics data. In order to differentiate between the 0 values representing the absence of the gene and 0 representing the non-expression of the gene, when compiling the data from different types of NSTI, we split the gene expression values in quartiles. Expression values for each gene above the second quartile (or median) are said to be expressed and the ones below the second quartile are said to be not expressed. The Human genes required no such treatment as the same genes were measured irrespective of NSTI type, hence the 0 value represented only non-expression.

4.2.5 Gaussian graphical models (GGM)

A Gaussian graphical model is a graphical representation of conditional dependencies of normally distributed variables. Gaussian graphical models assume that the variables in the data follow a multivariate Gaussian/normal distribution. The nodes in this type of graph represent the normally distributed variables and the presence or absence of the edges represents the conditional dependency or independency between the variables. We use the static shrinkage estimator described in **(Opgen-Rhein et al. 2006)** to build Gaussian graphical models between human and bacterial genes for stratified groups of patients based on the site of infection. We make the method more stringent by assigning a probability of conditional dependence to each edge of the graph and then pruning the edges based on a very high probability threshold (0.95). This is implemented based on the PCLRC algorithm tested for its robustness in **(Jahagirdar, Suarez-Diez, et al. 2019)**.

4.2.6 Mixed graphical models (MGM)

In this study, our interest lies in the association structures created by all the variables measured in NSTI patients as a proxy for understanding system-wide mechanisms and responses that govern the phenotypical observations. As detailed in Section

Gaussian graphical models (GGM), the Gaussian graphical model's reliance on the assumption of normal distribution implies that it cannot accommodate for any form of discrete variables. Many parameters recorded in the clinic are often discrete in nature with no possible continuous interpretation. Whole numbers representing for the antibiotic administered or the location of the infection have inherently no connotations for decimals/fractions. In order to calculate associations between continuous variables that can assume a normal distribution and discrete variables that can not, we follow the method introduced by **(J. Lee et al. 2013)** that combines the multivariate Gaussian and discrete pairwise Markov random field (also known as Ising model) to model the mixed variables as a pairwise Markov random field with an association that can be described as the joint probability distribution between the two variables. Since the number of variables ($m = p + q$) in our data far outnumber the number of samples (n), we utilise the pseudolikelihood method **(Besag 1975; J. Lee et al. 2013)** to predict the parameters determining the conditional distributions. We follow the example laid out by **(Altenbuchinger, Zacharias, et al. 2019)** in augmenting the negative pseudolikelihood with a LASSO penalty (λ) to calibrate overfitting. In this case, the pseudolikelihood was given by

$$\mathcal{L}(\Theta|x,y) = -\sum_{s=1}^{p} \log p(x_s|x_{\setminus s}, y; \Theta) - \sum_{r=1}^{q} \log p(y_r|x, y_{\setminus r}; \Theta), \tag{4.1}$$

where, $p(x_s|x_{\setminus s}, y; \Theta)$ is the conditional distribution of variable x_s that takes a Gaussian form and $p(y_r|x, y_{\setminus r}; \Theta)$ is the conditional distribution of a variable y_r that is discrete in nature with L_r states.

The conditional distributions can be defined as

$$p(x_s|x_{\setminus s}, y; \Theta) = \frac{\sqrt{-\beta_{ss}}}{\sqrt{2\pi}} \exp(a), \tag{4.2}$$

where,

$$a = \frac{\beta_{ss}}{2}\left(\frac{\alpha_s + \sum_j \rho_{sj}(y_j) + \sum_{t \neq s} \beta_{st} x_t}{\beta_{ss}} + x_s\right)^2. \tag{4.3}$$

$$p(y_r|x, y_{\setminus r}; \Theta) = \frac{\exp(b_{y_r})}{\sum_{l=1}^{L_r} \exp(b_l)}, \tag{4.4}$$

where,

$$b_{y_r} = \left(\sum_s \rho_{sr}(y_r) x_s + \frac{1}{2}\phi_{rr}(y_r, y_r) + \sum_{j \neq r} \phi_{rj}(y_r, y_j)\right), \tag{4.5}$$

and

$$b_l = \left(\sum_s \rho_{sr}(l) x_s + \frac{1}{2}\phi_{rr}(l, l) + \sum_{j \neq r} \phi_{rj}(l, y_j)\right). \tag{4.6}$$

Here, Θ is the inverse covariance, $x_{\setminus s}$ and y_r are the predictors for the linear regression determining the conditional distributions of x_s. x_s denotes the s^{th} of p continuous variables and y_j denotes the j^{th} of q discrete variables. The model parameters β_{st}, α_s,

$\rho_{sj}(y_j)$, and $\phi_{rj}(y_r, y_j)$ represent the continuous-continuous edge potential, continuous node potential, continuous-discrete edge potential, and discrete-discrete edge potential respectively. Here, potential represents a mathematical function that encodes the dependencies between variables. β_{st} is represented by a scalar where a 0 value equates to conditional independence between continuous variables x_s and x_t. Similarly ρ_{sj} is represented by a vector the size of L_j and a 0 value represents continuous independence between continuous variable x_s and discrete variable y_j. The penalty λ is integrated by

$$\text{minimise}_{\Theta} L_{\lambda}(\Theta) = L(\Theta) + \lambda \left(\sum_{s=2}^{p} \sum_{t=1}^{s-1} w_{st} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} w_{sj} \|\rho_{sj}\|_1 + \sum_{j=2}^{q} \sum_{r=1}^{j-1} w_{rj} \|\phi_{rj}\|_1 \right). \tag{4.7}$$

In this case, $\|.\|_1$ is defined as the element-wise matrix norm as shown in

$$\|\phi_{rj}\|_1 = \sum_{l=1}^{L_r} \sum_{m=1}^{L_j} |\phi_{rj}(l, m)|. \tag{4.8}$$

$$\|\rho_j\|_1 = \sum_{l=1}^{L_j} |\rho_j(l)|. \tag{4.9}$$

The weights w_{st}, w_{sj} and w_{rj} are given by the equations

$$w_{st} = \sigma_s \sigma_t. \tag{4.10}$$

$$w_{sj} = \sigma_s \sqrt{\sum_b q_b (1 - q_b)}. \tag{4.11}$$

$$w_{rj} = \sqrt{\sum_a p_a (1 - p_a) \sum_b q_b (1 - q_b)}. \tag{4.12}$$

Here σ_s is the standard deviation of the continuous variable x_s, $p_a = P_r(y_r = a)$, and $q_b = P_r(y_j = b)$.

We follow the example laid out in the supplementary of **(Altenbuchinger, Weihs, et al. 2020)** by determining the penalty (λ) based on the number of continuous variables (p), the number of discrete variables (q), and the number of samples (n). λ was determined by

$$\lambda = \delta \times \sqrt{\frac{log(p + q)}{n}}, \tag{4.13}$$

where δ was used as a multiplier and was set to 1 **(Altenbuchinger, Weihs, et al. 2020)**.

4.2.7 Random forest based conditional independence graphs

In the mixed graphical model applied above, the mixing of two different types of models can come with its own inherent limitations such as when predicting for a discrete variable, each level contributes to an additive effect whereas predicting a continuous variable contributes linear effects **(J. Lee et al. 2013)**. Random Forests is a method that inherently can deal with these mixed-type variables **(Breiman 2001)**. Random Forests is an ensemble machine learning method based on decision trees. Furthermore, a sub-sampling-based probability (p-values) threshold can be used on the random forest feature importance to control the stringency of the model result **(Archer et al. 2016)**. We use the method proposed by **(Fellinghauer et al. 2013)** to build conditional independence/dependence graphs using the random forest algorithm with stability selection based on a similar subsampling approach. In this GRaFo **(Fellinghauer et al. 2013)** method approach, the edge $i-j$ association is obtained from the feature importance. The ranking of edges of mixed variables is based on p-values and a stability selection procedure. The stability selection procedure is based on an upper bound of false positives/type I errors ($\mathbb{E}[V]$) in order to impose a stringent threshold on the edges **(Fellinghauer et al. 2013; Meinshausen et al. 2010)**. Only the edges $i-j$ are chosen that fulfil the criterion of

$$\frac{1}{n_{sub}}\sum_{k=1}^{n_{sub}} I_{\{i-j\in\varepsilon(\hat{G}_{CIG}(X^{(k)}))\}} \geq \pi_{thres}. \tag{4.14}$$

where, $\varepsilon(\hat{G}_{CIG}(X^{(k)}))$ represents the edges from a threshold ranking based on $X^{(k)}$, π_{thres} is a threshold on the minimum relative frequency of edges across the n_{sub} subsets, and I is the indicator function. The π_{thres} takes values $\pi_{thres}\in\left(\frac{1}{2},1\right)$ and is determined based on the acceptable number of false positives ($\mathbb{E}[V]$) such that,

$$\mathbb{E}[V] \leq \frac{q^2}{(2\pi_{thres}-1).m.\frac{(m-1)}{2}}. \tag{4.15}$$

where q represents the number of selected edges per subset and $m.(m-1)/2$ is the total number of possible edges. We run this method with 10000 decision trees for each prediction, $n_{sub}=100$, and accepting only 5% of all edges as false positives such that $\mathbb{E}[V]=0.05\times m.(m-1)/2$. This application of random forest also allows us to run the random forest algorithm predicting certain variables of interest with a greater number of decision trees and more stringent probability-based ranking of the feature importance without building a conditional independence/dependence network to understand the associations to the variable of importance. We run a few predictions for variables of importance with 100000 decision trees and 100 repetitions with subsampling.

4.2.8 Interpreting networks

The networks from the GGM can be interpreted as a pair-wise partial correlation where a 0 value represents a conditional Independence and the conditional dependence takes the values $\in(-1,1)$. The positive values represent a positive association

where an increase and decrease in the level of one variable corresponds to an increase or decrease in the values of the other. The negative values represent a negative association where the increase in the level of one variable corresponds to a decrease in the values of the other variable and vice versa.

The networks from the MGM can be interpreted as a pairwise conditional dependence where a 0 value once again represents conditional independence and the edges take the values $\in (-\infty, \infty)$.

The networks from the Random Forest based GRaFo method can be interpreted as conditional independence graphs where the edges take the binary values of 0 and 1. 0 represents conditional independence and 1 represents conditional dependence.

4.2.9 Network topology: hubs and motifs

Throughout this chapter, we refer to conditional dependencies/connections with designations of primary or secondary conditional dependencies. This is in reference to the positions of the two nodes in the network relative to each other. When the two variables are conditionally dependent on each other, it could be characterised by the designation primary. However, if the parameters are conditionally dependent with a third parameter, then they can be characterised with the designation of secondary. We present a modified figure from **(Jahagirdar, Suarez-Diez, et al. 2019)** as an example in figure 4.1 to showcase the designations. In order to understand and study the topology of the big network, we focus on the network hubs and network motifs. We label a node/variable as a hub based on the number of conditional dependencies that the node has with other nodes/variables. We set a minimum threshold as 5 conditional dependencies for a node to be labelled as a hub and the importance of the hub was assumed to be directly proportional to the number of conditional dependencies of the said hub. Here, we also define three-node motifs as a network structure formed when 3 nodes have conditional dependencies with each other. We add a further condition that at least one of the nodes needs to represent a clinical variable.

4.2.10 Software

R and Python were used for all of the analyses. Data transformations were done in R **(Team et al. 2016)**. The missforest package was used for the missing data in clinical parameters **(Stekhoven et al. 2012)**. The caret package was used to create dummy variables and handle variable transformations **(Kuhn et al. 2020)**. The GeneNet package was used for GGM models **(J. Schäfer, Opgen-Rhein, et al. 2006)**. The R and python codes for the MGM implementation were used from **(Altenbuchinger, Zacharias, et al. 2019; Altenbuchinger, Weihs, et al. 2020)**. The R codes for the GRaFo implementation were taken from **(Fellinghauer et al. 2013)** and the rfPermute package was used for p-value calculations on the random forest importance **(Archer et al. 2016)**. The Gene ontology (GO) enrichment analysis was performed using the PANTHER tool **(Mi et al. 2019)** and the pathway enrichment analysis was done using the tools of Reactome **(M. Gillespie et al. 2022; Griss et al. 2020)**. The bacterial genes were proteins and functions were derived from UniProt and UniParc **(T. U. Consortium 2023)**. All the codes for detecting hubs and motifs were written

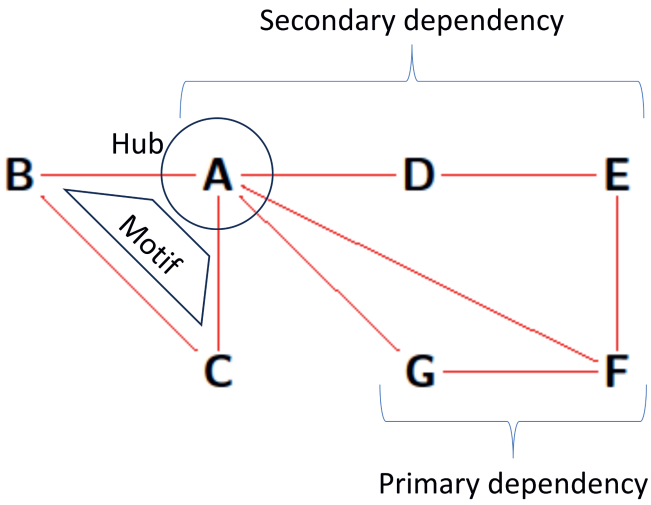


Figure 4.1: This example was modified from **(Jahagirdar, Suarez-Diez, et al. 2019)** to demonstrate the designations of hubs, motifs, and primary and secondary connections. In this example, a toy network is shown where, A,B,C,D,E,F, & G are nodes representing parameters and the red lines between them represent the conditional dependencies. G & F can be said to be primary connections as they are conditionally dependent on each other. A & E can be said to be secondary connections as they are conditionally dependent on D. This logic can be further extended ad infinitum. A can be said to be a network hub as it has significantly more conditional dependencies than other variables. A, B & C is said to have formed a network motif as A & B are conditionally dependent, B & C are conditionally dependent, and C & A are conditionally dependent if one of the three nodes is a clinical variable.

in R and the networks were visualised using the anvis package from the chapter *Automated NETwork VISualisation with anvis*. The supplementary files are available on Zenodo: https://doi.org/10.5281/zenodo.8128358

4.3 Results and Discussion

A total of 2240 variables were included in the analysis consisting of 676 clinical variables, 680 human genes, 845 bacterial genes, and 39 plasma analytes. 44 samples were included in the study to construct the models. Dummy parameters were constructed out of clinical variables with discrete measurements making the final network consisting of 2725 parameters and 6462178 conditional dependencies.

4.3.1 The NSTI multi-omics network

The mixed graphical model constructed from NSTI patient samples depicts a system-wide response manifested as an association structure. This response reflects the interplay between bacteria and host in NSTI patients, as captured by multi-omics measurements. By adopting a system-wide approach, we anticipate that this study will

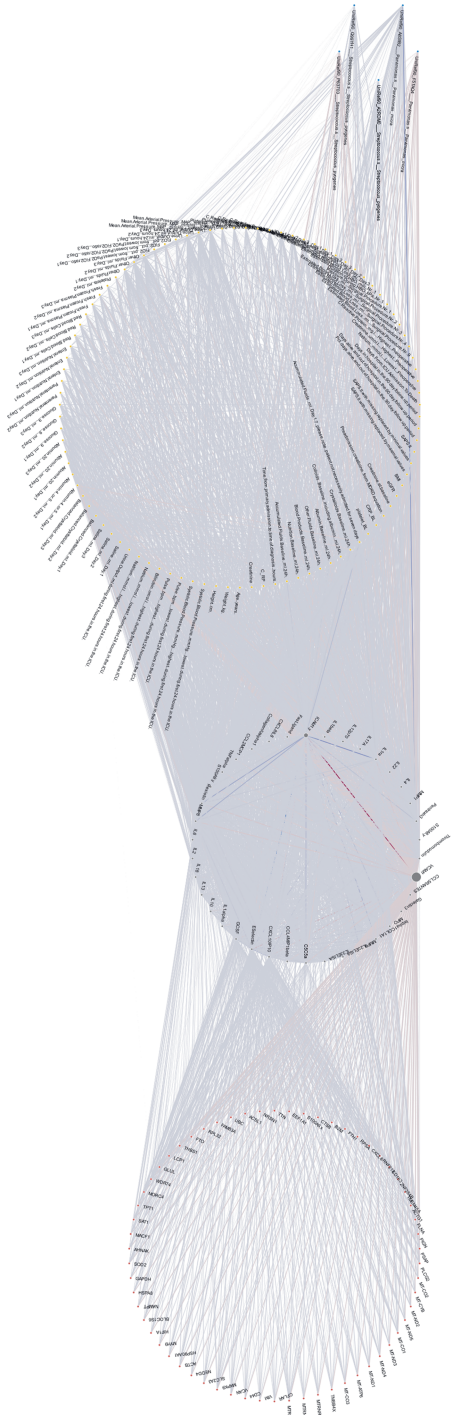


Figure 4.2: The top 5% conditional dependencies of the network created using the mixed graphical model.

Figure 4.2: The parameters are separated into 4 circles representing (from top to bottom) bacterial genes, clinical parameters, human genes and plasma analytes. The red lines connecting the parameters are strong positive conditional dependencies and the blue line connecting the parameters are strong negative conditional dependencies. The bacterial genes are UniRef90 Ids with a prefix of Sp or Pm referring to the gene belonging to *S. pyogenes* or *P. micra* respectively.

reveal markers of general cellular mechanisms involved in growth, proliferation, and differentiation alongside those associated with the spread of infection. The top 5% edges of the mixed graphical model are shown in 4.2 where the 4 circles from top to bottom represent bacterial genes, clinical variables, human genes and plasma analytes. The genes and plasma analytes from this study have been studied in-depth elsewhere **(Medina et al. 2021; Rath et al. 2023; Jahagirdar, L. Morris, et al. 2022; Thänert et al. 2019)**. In this study, we focus on the associations between these various multi-omics measurements and particularly associations with variables measured in the clinic. Hence we focus more on 4.3 which shows the conditional dependencies between different clinical variables, genes, and analytes.

Days from ICU admission to death is the central node of the network with a strong positive conditional dependence with the protein VCAM-1, a member of the immunoglobulin family. Days from ICU admission to death also has strong negative conditional dependence with the human gene WDR74 and several bacterial genes from *S. pyogenes* and *P. micra* including genes encoding for LacD1 and LacA1, known virulence factors in forming a bidirectional link between *S. pyogenes* metabolism and regulation of pathogenesis **(Pancholi et al. 2022)**.

4.3.2 Network topology

In order to conduct a thorough data-driven analysis of the network from the mixed graphical model, we studied the topology of the network by identifying network hubs and motifs. Network hubs are nodes in the network with a significantly greater number of edges/connections than the average number of connections per node as elaborated in *Network topology: hubs and motifs*. In this case, nodes are the variables, and connections are the conditional dependencies. The top 25 network hubs for each data type are shown in Table 4.1 along with the number of conditional dependencies the hub possesses.

Even though the plasma analytes have far more conditional dependencies, the conditional dependencies themselves have lower values in general compared to the clinical parameters. The complete network can be accessed in supplementary file S1 Chapter4. Most of the important clinical variables based on the number of conditional dependencies are variables registered daily in the ICU.

Many human genes deemed significant here were also found to have significant interactions with bacterial genes in another study analysing the host-pathogen interactions **(Jahagirdar, L. Morris, et al. 2022)** in **Chapter 5**. The genes MT-RNR2 Like 12 pseudogene (MTRNR2L12), Neurensin 1 (NRSN1), Zinc Finger Protein 354B (ZNF354B), Ferritin Heavy Chain 1 (FTH1), NADH dehydrogenase subunit 4 (MT-ND4L), and Mitochondrially encoded NADH: ubiquinone oxidoreductase core subunit 1 (MT-ND1) were all found to have significant interactions with bacterial genes

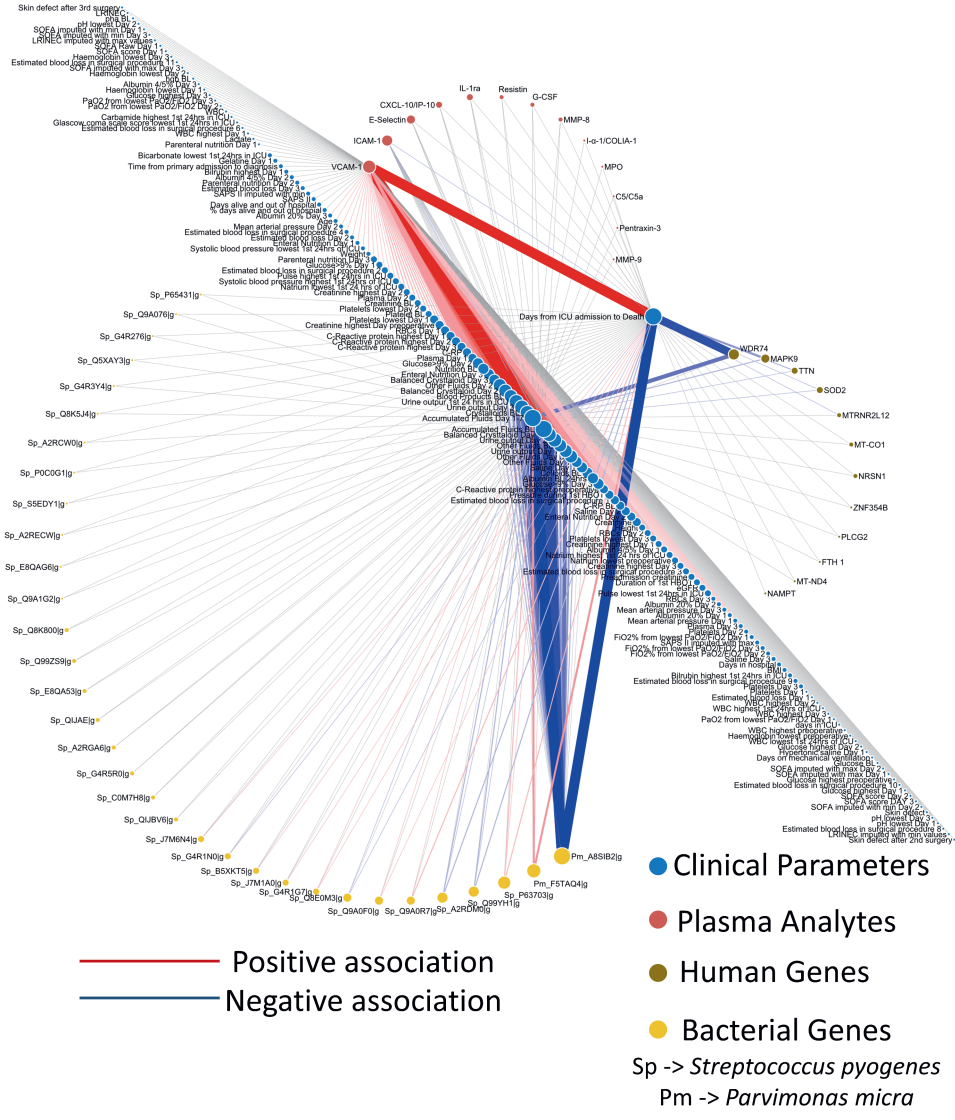

IL-1ra
Resistin
G-CSF
CXCL-10/IP-10
E-Selectin
MMP-8
ICAM-1
α-1/COLIA-1
VCAM-1
MPO
C5/C5a
Pentraxin-3
MMP-9
Days from ICU admission to Death
WDR74
MAPK9
TTN
SOD2
MTRNR2L12
MT-CO1
NRSN1
ZNF354B
PLCG2
FTH 1
MT-ND4
NAMPT
Sp_P65431|g
Sp_Q9A076|g
Sp_G4R276|g
Sp_Q5XAY3|g
Sp_G4R3Y4|g
Sp_Q8K5J4|g
Sp_A2RCW0|g
Sp_P0C0G1|g
Sp_S5EDY1|g
Sp_A2RECW|g
Sp_E8QAG6|g
Sp_Q9A1G2|g
Sp_Q8K800|g
Sp_Q99ZS9|g
Sp_E8QA53|g
Sp_QIJAE|g
Sp_A2RGA6|g
Sp_G4R5R0|g
Sp_C0M7H8|g
Sp_QIJBV6|g
Sp_J7M6N4|g
Sp_G4R1N0|g
Sp_B5XKT5|g
Sp_J7M1A0|g
Sp_G4R1G7|g
Sp_Q8E0M3|g
Sp_Q9A0F0|g
Sp_Q9A0R7|g
Sp_A2RDM0|g
Sp_Q99YH1|g
Sp_P83703|g
Pm_F5TAQ4|g
Pm_A8SIB2|g
Positive association
Negative association
Clinical Parameters
Plasma Analytes
Human Genes
Bacterial Genes
Sp -> Streptococcus pyogenes
Pm -> Parvimonas micra

Figure 4.3

Table 4.1: The table highlights the most important hubs in the network created with the mixed graphical model. The table is divided into 4 sections. The 4 sections isolate and show the biggest network hubs from clinical parameters, human genes, bacterial genes and plasma analytes. Each of these 4 sections are organised vertically in a descending order of the number connected nodes each parameter has. Thus, the parameters in clinical parameters can be read as the biggest hubs that are clinical parameters in the network. Similarly, for human genes, they are the biggest hubs that are human genes and so on and so forth. There is no horizntal organisation between the various groups in this table. The table shows the parameters and the conditional dependencies the parameter has in the network. In the case of human and bacterial genes, the table also summarises either the associated protein or function of the gene. The bacterial genes are UniRef90 Ids with a prefix of Sp or Pm referring to the gene belonging to *S. pyogenes* or *P. micra* respectively.

Clinical Parameters	Connected Nodes	Human Genes	Associated Protein/function	Connected Nodes	Bacterial Genes	Associated protein/function	Connected Nodes	Plasma Analytes	Connected Nodes
Days from ICU admission to death	136	WDR74	WD Repeat Domain 74	15	Pm-A8SIB2	Uncharacterized protein	52	VCAM-1	1995
Accumulated fluids Day1-7	135	MAPK9	Mitogen-Activated Protein Kinase 9	11	Pm-F5TAQ4	Glycine/sarcosine/b reductase complex protein A	21	ICAM-1	1003
Accumulated Fluids BL (24hrs)	91	SOD2	Superoxide Dismutase 2	10	Sp-P63703	lacD1(Tagatose 1,6-diphosphate aldolase 1)	20	E-Selectin	931
Crystalloids BL (24hrs)	65	MTRNR2L12	MT-RNR2 Like 12 (Pseudogene)	10	Sp-Q99YH1	lacA1(Galactose-6-phosphate isomerase subunit LacA 1)	20	CXCL-10/IP-10	610
Balanced crysttaloid Day1	45	MT-CO1	Mitochondrially Encoded Cytochrome C Oxidase I	10	Sp-A2RDM0	Hypothetical cytosolic protein	20	IL-1RA	355
Urine output (24hrs) Day3	44	TTN	Titin	11	Sp-Q9A0R7	Putative gluconeogenesis factor	18	G-CSF	201
Other fluids BL (24hrs)	44	NRSN1	Neurensin 1	10	Sp-Q9A0F0	cmk(Cytidylate kinase)	18	Resistin	193
Blood products BL (24hrs)	40	ZNF354B	Zinc Finger Protein 354B	10	Sp-Q8E0M3	Catabolite control protein A	18	MMP-8	182
Saline Day1	27	PLCG2	Phospholipase C Gamma 2	10	Sp-G4R1G7	Uncharacterized protein	18	MPO	170
Enteral nutrition Day3	24	FTH1	Ferritin Heavy Chain 1	8	Sp-J7M1A0	Phosphohydrolase	18	I-α-1/COL1A1	172
Colloids BL (including albumin) (24hrs)	22	MT-ND4	Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 4	7	Sp-B5XKT5	Ferredoxin	18	C5/C5a	150
Nutrition BL (24hrs)	22	NAMPT	Nicotinamide Phosphoribosyltransferase	7	Sp-G4R1N0	Uncharacterized protein	18	MMP-9	123
Albumin BL (24hrs)	20	ACTB	Actin Beta	7	Sp-J7M6N4	pepA(Glutamyl aminopeptidase)	18	Pentraxin-3	132
Glucose Day2	21	EEF1A1	Eukaryotic Translation Elongation Factor 1 Alpha 1	7	Sp-Q1JBV6	potA(Spermidine/pu import ATP-binding protein PotA)	18	Thrombomodulin	107
Highest C-RP Preoperative	20	AHNAK	AHNAK Nucleoprotein	7	Sp-C0M7H8	ldh(L-lactate dehydrogenase)	18	IL-6	112
C-RP	20	MT-CO3	Mitochondrially Encoded Cytochrome C Oxidase III	6	Sp-G4R5R0	SAM-dependent methyltransferase	18	CCL-5/RANTES	70
Pressure During HBO (Nr 1)	20	UBC	Ubiquitin C	5	Sp-A2RGA6	Acetyltransferase (GNAT) family	17	S100A9	50
Estimated Blood Loss (Surgical Procedure Nr 1)	20	MT-ND5	Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 5	5	Sp-Q1JAE1	Uncharacterized protein	17	IL-23	44
Red Blood Cells Day1	18	B2M	Beta-2-Microglobulin	5	Sp-E8QA53	proV(Glycine betaine transport ATP-binding protein)	17	MMP-1	46
Highest Creatinine (Preoperative)	18	MT-ND1	Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 1	5	Sp-Q99ZS9	murB(UDP-N-acetylenolpyruvoylgl reductase)	17	Galectin-3	46
Lowest Platelets Day1	18	MT-ATP6	Mitochondrially Encoded ATP Synthase Membrane Subunit 6	5	Sp-Q8K800	Membrane protein	17	Collagen-IVα1	45
Height	18	MT-ND2	Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 2	5	Sp-Q9A1G2	iscU(Iron-sulfur cluster assembly scaffold protein IscU)	16	IL33	45
Creatinine BL	18	SAT1	Spermidine/Spermin N1-Acetyltransferase 1	5	Sd-E8QAG6	Protein essC	16	CCL-2/MCP-1	44
Albumin Day1	20	GLUL	Glutamate-Ammonia Ligase	5	Sp-A2REC2	Protein ADP-ribosyltransferase	16	IL-13	40
Highest Natrium (1st 24hrs in ICU)	17	TPT1	Tumor Protein, Translationally-Controlled 1	5	Sp-P0C0G1	folP(Dihydropteroat synthase)	16	S100A8	39

Figure 4.3: The conditional dependencies between parameters from different data sets. In this figure, we are not interested in the conditional dependencies between two clinical parameters, but instead, we are interested in the conditional dependencies between a clinical parameter and a bacterial gene or a human gene. The nodes on the diagonal line, coloured in blue, represent the clinical parameters. The nodes on the left in a semicircle, coloured in yellow represent bacterial genes. The nodes on the top left, coloured in red represent plasma analytes, and the nodes on the right, coloured in olive represent human genes. The size of the nodes is directly proportional to the number and magnitude of conditional dependencies that parameter makes with the other parameters in the network. The conditional dependencies are represented by the lines/edges connecting these parameters. The width of these lines is exponentially proportional to the magnitude of the conditional dependency. The red colour in the lines represent a strong positive conditional dependency. the blue lines in the network represent strong negative conditional dependency. The bacterial genes isolated in the figure are from two different species. The genes from *S. pyogenes* are designated by the prefix Sp and the genes from *P. micra* are designated with the prefix Pm. The bacterial genes are UniRef90 Ids.

in the host-pathogen analysis **(Jahagirdar, L. Morris, et al. 2022)**. Upon conducting Gene Ontology (GO) enrichment analysis on human genes, we discovered several significantly enriched GO terms that describe the biological processes and molecular functions involved. Specifically, our analysis revealed that the following GO terms were enriched for complete biological process: electron transport coupled proton transport, mitochondrial electron transport, NADH to ubiquinone, response to metal ion, and proton motive force-driven mitochondrial ATP synthesis. Additionally, the following GO terms were enriched for molecular function: diamine N-acetyltransferase, NADH dehydrogenase, and electron transfer activity. When performing pathway enrichment analysis, we find the pathways associated with tRNA and rRNA processing, DNA damage recognition, several pathways relating to signal transduction and immune system including JNK phosphorylation, MAP kinase activation, regulation of NF-$\kappa\beta$ signalling, and NRIF signals enriched. Complete analyses can be found in supplementary file S1 Chapter4.

The important bacterial genes consist of known virulence factors, essential genes (MurB), protein transporters (PotA, PotV), drug targets (Ferredoxin), and 1 gene recognised for antibiotic resistance (FolP). The genes encoding for LacD1 (P63703), LacA1 (Q99YH1), and catabolic control protein A (Q8E0M3) are recognised virulence factors in the PATRIC database **(Wattam et al. 2017)**. Carbon catabolite repression has been known to affect the expression of virulence determinants through catabolic control protein A (Q8E0M3) and LacD1 (P63703), where they act as carbohydrate-sensitive regulators **(L. A. Vega et al. 2022)**. LacD1 (P63703) has also been recognised as a global regulator of virulence factor expression in *S. pyogenes* **(Loughman et al. 2006)**. LacA1 (Q99YH1) is involved in the tagatose pathway and has a role to play in *S. pyogenes* switching its carbon metabolism from glucose-based to favouring galactose influenced by Zn^{2+} concentration when *S. pyogenes* is in contact with neutrophils **(Pancholi et al. 2022)**.

The top plasma analytes are the three soluble adhesion molecules which are cell surface binding proteins. ICAM-1 and VCAM-1 belong to the immunoglobulin domain and are important for firm adhesion and transendothelial migration whereas E-

Selectin belongs to the selectin family and has functions related to leukocyte rolling. Leukocyte transendothelial migration is a key step in the recruitment of leukocytes like neutrophils to sites of inflammation, injury, and immune reactions **(Yang, Froio, et al. 2005)**. This recruitment involves a multi-step cascade consisting of leukocyte rolling, firm adhesion, and transmigration **(Springer 1994; Yang, Froio, et al. 2005)**. VCAM-1, ICAM-1, and E-Selectin all play a crucial role in the migration and adhesion of leukocytes such as neutrophils **(Yang, Froio, et al. 2005; Rosner et al. 2001)**. Overall they are markers for inflammatory processes related to activation and damage to platelets and the endothelium **(Zonneveld et al. 2014)**.

We look for motifs in the network consisting of 3 nodes/variables where at least one variable is a clinical parameter due to our interest in the associations with the clinical variables. These network motifs could be considered as a unit with potential functional properties **(Shen-Orr et al. 2002)**. A select number of motifs are shown in Figure 4.4. The 3 conditional dependencies between the 3 variables could have a positive or negative association. When all three conditional dependencies are positive, we could interpret it as an indirect positive feedback loop. Such a positive feedback loop can be observed in the mixed graphical model between the parameters Accumulated Fluids baseline, VCAM-1, and LacD1 (P63703). Similarly, when all three conditional dependencies between the 3 variables are negative, it could be interpreted as an indirect negative feedback loop. Such a negative feedback loop can be observed in the mixed graphical model between the parameters Urine output Day 1, Thrombomodulin, and MMP-8.

4.3.3 Hypotheses of interest

One of the aims of this exploratory analysis is to study all the parameters from a systems medicine perspective and look for associations with variables of particular interest from other studies **(Jahagirdar, L. Morris, et al. 2022; Thänert et al. 2019; Medina et al. 2021; Rath et al. 2023)** to understand the underlying mechanisms and generate hypotheses that can be further pursued. In this study, we employ the GRaFo model to construct highly stringent conditional independence graphs. These graphs enable us to investigate associations with parameters deemed crucial within the network (as identified in sections *The NSTI multi-omics network* and *Network topology*), as well as with predetermined parameters explored during the study's design phase.

We explore the primary and secondary conditional dependencies of Thrombomodulin, C-reactive protein, WDR74, TTN, Urine output Day 3, and *S. pyogenes* LacA1 (Q99YH1) in Figure 4.5A, B, C, D, E, and F respectively. We further explore the associations with Thrombomodulin, Mortality, and microbial aetiology with respect to their significance before the threshold is applied by the GRaFo method. The C-reactive protein measurements had conditional dependencies on other C-reactive protein related measurements and the plasma analyte C5/C5a which is involved in the complement cascade. The GO enrichment analysis for the associations of the gene WRD74 reveals the GO term 3-O-(N-acetyl-D-glucosaminyl)-L-serine O-N-acetyl-alpha-D-glucosaminase activity in terms of molecular function and the term activation of store-operated calcium channel activity in terms of biological process. The pathway enrichment analysis highlights pathways of cytokine signalling in immune system

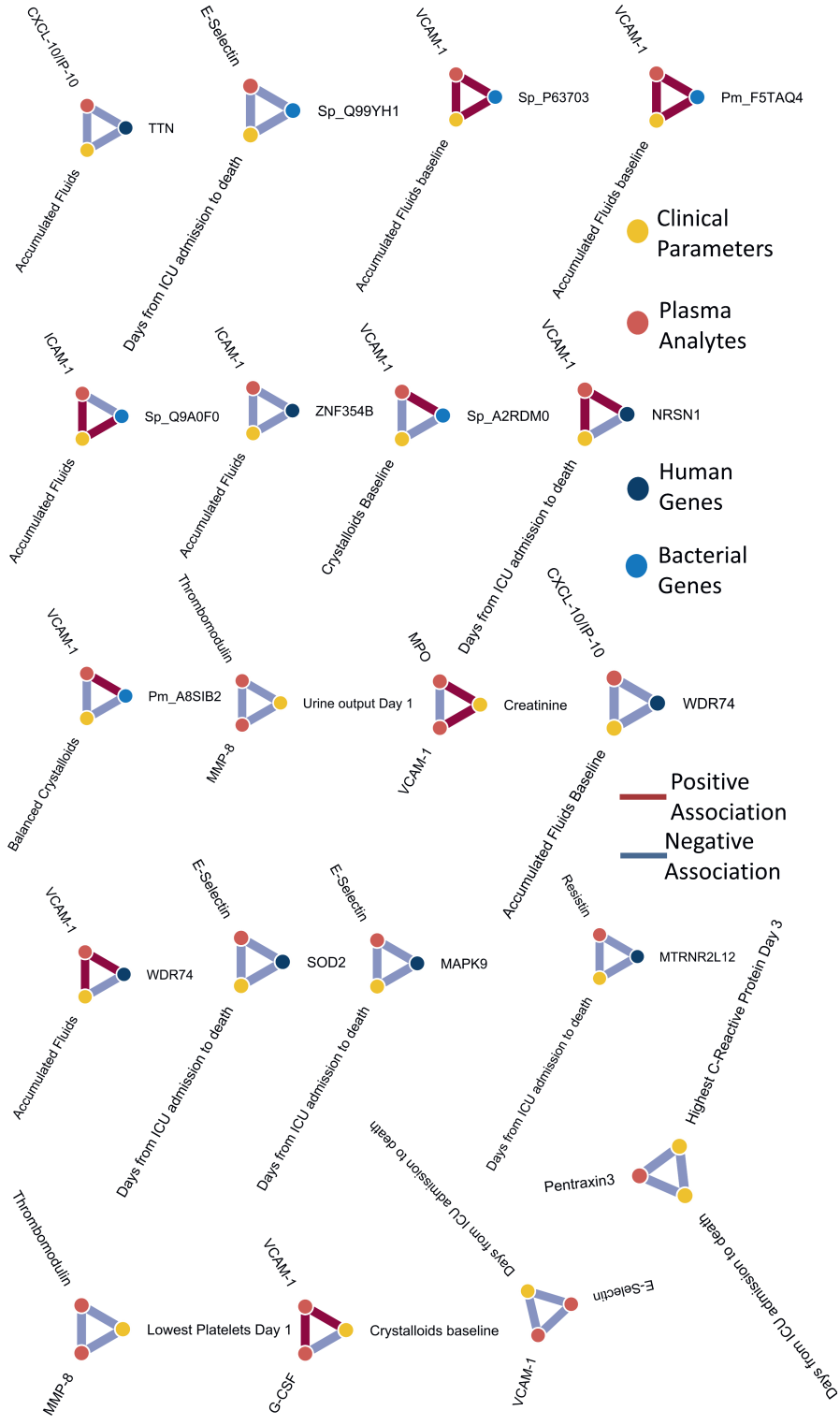

CXCL-10/IP-10
TTN
Accumulated Fluids
E-Selectin
Sp_Q99YH1
Days from ICU admission to death
VCAM-1
Sp_P63703
Accumulated Fluids baseline
VCAM-1
Pm_F5TAQ4
Accumulated Fluids baseline
Clinical Parameters
Plasma Analytes
ICAM-1
Sp_Q9A0F0
Accumulated Fluids
ICAM-1
ZNF354B
Accumulated Fluids
VCAM-1
Sp_A2RDM0
Crystalloids Baseline
VCAM-1
NRSN1
Days from ICU admission to death
Human Genes
Bacterial Genes
VCAM-1
Pm_A8SIB2
Balanced Crystalloids
Thrombomodulin
Urine output Day 1
MMP-8
MPO
Creatinine
VCAM-1
CXCL-10/IP-10
WDR74
Days from ICU admission to death
Positive Association
Negative Association
VCAM-1
WDR74
Accumulated Fluids
E-Selectin
SOD2
Days from ICU admission to death
E-Selectin
MAPK9
Days from ICU admission to death
Accumulated Fluids Baseline
Resistin
MTRNR2L12
Days from ICU admission to death
Highest C-Reactive Protein Day 3
Pentraxin3
Days from ICU admission to death
Thrombomodulin
Lowest Platelets Day 1
MMP-8
VCAM-1
Crystalloids baseline
G-CSF
Days from ICU admission to death
E-Selectin
VCAM-1

Figure 4.4

Figure 4.4: Network motifs are repeating topological structures in a network to which a function could be assigned. This figure shows a select number of network motifs consisting of 3 parameters each. At least one parameter in each motif is a clinical parameter. The clinical parameters, plasma analytes, human genes and bacterial genes are denoted with the colours yellow, red, dark blue and light blue respectively. The conditional dependencies are red in colour if they are positive and blue in colour if they are negative. The bacterial genes are UniRef90 Ids with a prefix of Sp or Pm referring to the gene belonging to *S. pyogenes* or *P. micra* respectively.

and signal transduction including signalling by CSF1 in myeloid cells, IL-6-type cytokine receptor ligand interactions, IL-10 signalling, IL-20 signalling, other interleukin signalling, FCERI mediated MAPK activation, and EGFR Transactivation by Gastrin. The analysis also highlighted the enrichment of potential therapeutics for SARS.

The GO enrichment analysis for the associations with the gene TTN reveals the GO terms cardiac muscle thin filament assembly and skeletal muscle thin filament assembly in biological processes and the terms muscle alpha-actin binding and titin binding in molecular functions. The pathway enrichment analysis shows the pathway striated muscle contraction to be enriched. Many of the primary and secondary associations with urine output are measurements taken in the ICU along with scores accessing the performance of several organ systems such as SOFA and SAPS II. The primary and secondary associations with LacA1 (Q99YH1) reveal associations with virulence factors, essential genes and transporter proteins. Pyruvate formate lyase activating enzyme (M4YVM2), Transcriptional regulator, DeOR family(K4Q6Q7), lead, cadmium, zinc & mercury transporting ATPase (J7M3S0), and 3-dehydroquinate dehydratase (H8HCL1) are all recognised virulence factors in the PATRIC database **(Wattam et al. 2017)**. Nicotinate phosphoribosyltransferase (F7IV70) and Glutamyl-tRNA amidotransferase subunit C (P68890) are known to be essential genes while the putative membrane protein (U2USH5) is a transport protein.

Thrombomodulin and differences in coagulation in NSTI patients

Thrombomodulin is an integral membrane protein that inhibits coagulation by forming the thrombin-thrombomodulin complex and activating protein C **(Levi and Poll 2017)**. Recombinant Thrombomodulin has been shown to reduce 28-day and in-hospital mortality of a subset of severe sepsis patients with severe coagulopathy, high fibrinogen/fibrin-degradation-products, D-dimer levels, severe organ dysfunction, and high mortality **(Kudo et al. 2021)**. The clinical variables, human and bacterial genes, and the plasma analytes associated with Thrombomodulin along with their significance are shown in table 4.2.

Thrombomodulin level is associated with creatinine, lactate, platelets, and BMI measures along with SOFA scores. Creatinine, lactate, platelets, and SOFA scores are also indicators of sepsis severity **(FRooN et al. 1994; S. M. Lee et al. 2016; Guclu et al. 2013; C. Liu et al. 2022)**. Although the implication of the association with BMI is not exactly clear, a higher BMI potentially results in a larger surgical area. We find that the conditional dependence between Thrombomodulin and body surface area is also significant (not shown here, see supplementary file S1 Chapter 4). The Human

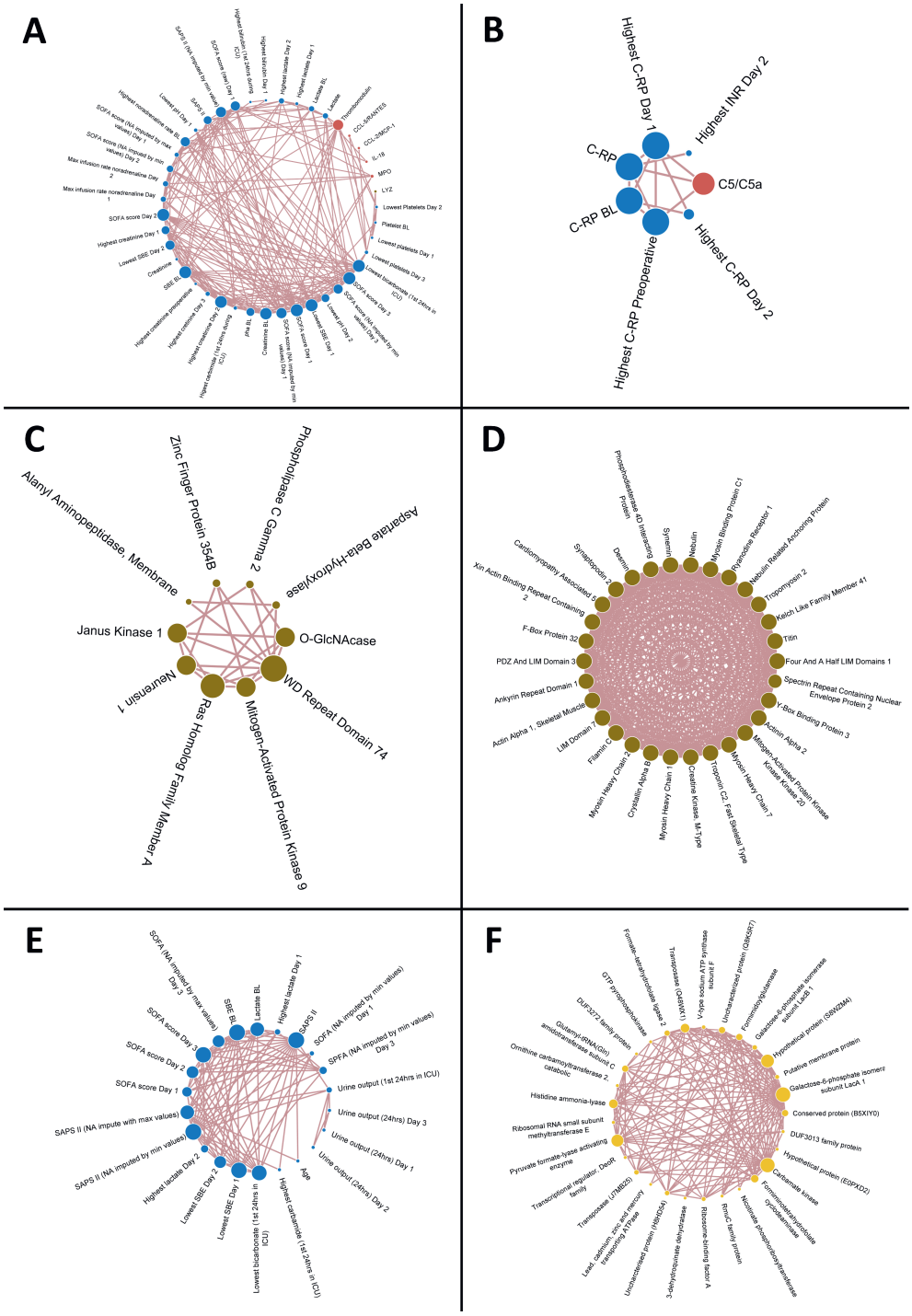

A
B
Highest C-RP Day 1
Highest INR Day 2
C-RP
C5/C5a
C-RP BL
Highest C-RP Preoperative
Highest C-RP Day 2
C
Zinc Finger Protein 354B
Phospholipase C Gamma 2
Alanyl Aminopeptidase, Membrane
Aspartate Beta-Hydroxylase
Janus Kinase 1
O-GlcNAcase
Neurensin 1
WD Repeat Domain 74
Ras Homolog Family Member A
Mitogen-Activated Protein Kinase 9
D
Phosphodiesterase 4D Interacting Protein
Synemin
Nebulin
Myosin Binding Protein C1
Synaptopodin 2
Desmin
Nebulin Related Anchoring Protein
Cardiomyopathy Associated 5
Tropomyosin 2
Xin Actin Binding Repeat Containing 2
Kelch Like Family Member 41
F-Box Protein 32
Titin
PDZ And LIM Domain 3
Four And A Half LIM Domains 1
Ankyrin Repeat Domain 1
Spectrin Repeat Containing Nuclear Envelope Protein 2
Actin Alpha 1, Skeletal Muscle
Y-Box Binding Protein 3
LIM Domain 7
Actinin Alpha 2
Filamin C
Mitogen-Activated Protein Kinase Kinase Kinase 20
Myosin Heavy Chain 2
Troponin C2, Fast Skeletal Type
Crystallin Alpha B
Myosin Heavy Chain 7
Myosin Heavy Chain 1
Creatine Kinase, M-Type
E
F

Figure 4.5

Figure 4.5: The sub-graphs from the GRaFo model are shown by isolating the primary and secondary conditional dependencies of a parameter of interest. The nodes are coloured based on the scheme seen in figure 4.3 such that the clinical parameters, bacterial genes, human genes and plasma analytes are coloured blue, yellow, olive, and red respectively. The red lines between parameters shows the existence of a conditional dependence and the lack of a line between parameters shows the existence of conditional independence. **(A)** The primary and secondary conditional dependencies of Thrombomodulin. **(B)** The primary and secondary conditional dependencies of the measurement of C-Reactive protein. **(C)** The primary and secondary conditional dependencies of the human gene WDR74. **(D)** The primary and secondary conditional dependencies of the human gene TTN. Figure **(E)** The primary and secondary conditional dependencies of urine output on day 3 of admission. **(F)** The primary and secondary conditional dependencies of *S. pyogenes* gene lacA1 (Q99YH1). The bacterial genes are givien in UniRef90 Ids.

genes result in the enrichment of the GO terms endothelial cell-matrix adhesion, induction of programmed cell death, platelet-derived growth factor receptor-alpha signalling pathway, and negative regulation of plasma cell differentiation in biological processes in terms of biological processes and the GO terms platelet-derived growth factor alpha-receptor activity, complement component C4b receptor activity, and interleukin-1 type II receptor antagonist activity in terms of molecular function. The pathway enrichment analysis reveals Interleukin-10 signalling, NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux, RHO GTPase cycle, Defective ABCA1 causes TGD, Signalling by PDGFRA extracellular mutants, and Ciprofloxacin ADME expressed. The complete results can be found in supplementary file S?. The functions of the bacterial genes associated with Thrombomodulin fall into the categories of virulence factors, essential genes, transporter proteins, and antibiotic resistance. Lmb (Q9ZHG8), Putative cell-cycle regulation histidine triad protein (A2RCZ2), pneumococcal histidine triad protein (J7M9C9), M protein (JZM209), and Trans-acting positive regulator (J7M9C1) are all recognised virulence factors on the PATRIC database. 50S ribosomal protein L16 (C0M8M3), Adenylate kynase (Q3K3U8), and 30S ribosomal protein S17 (Q3K3W0) are recognised as essential genes. Nucleaside permease (J7M618) is a transporter protein and 30S ribosomal protein S10 has a function in antibiotic resistance.

HrcA (Q5XAD4) has been shown to be a negative regulator of principle bacterial heat shock proteins (DnaK and GroES) that act as the cell's principal molecular chaperones responding to environmental stress **(Woodbury et al. 2003)**. DnaK (U9WVJ8) is also found to have conditional dependence with Thrombomodulin as seen in table 4.2. The heat shock response proteins HrcA (Q5XAD4) and DnaK (U9WVJ8) are also vital in the bacterial response to increased body temperature and fever **(Yura et al. 1999)**.

The immunity factor for SPN (H8HBV4) is a protein that forms an inhibitory complex by binding to the toxin *S. pyogenes* $\beta - NAD^+$ glycohydrolase (SPN) in order to keep the bacteria immune from the effects of the SPN-toxin **(C. L. Smith et al. 2011)**. SPN found in the Streptolysin O-operon, is one of the most potent toxins in *S. pyogenes* **(C. L. Smith et al. 2011)**. The toxin NAD^+-glycohydrolase has been shown to promote the intracellular survival of *S. pyogenes* **(Sharma et al. 2016)**.

HrcA (Q5XAD4) has been implicated in the formation of biofilm in several or-

ganisms such as *Listeria monocytogenes* and *Helicobacter pylori* **(Veen et al. 2010; Hathroubi et al. 2018)**. This biofilm formation has also been previously documented for *S. pyogenes* during NSTI **(Siemens, Chakrakodi, et al. 2016)**. The toxin NAD^+-glycohydrolase has been shown to be essential for the optimal formation of such biofilms by *S. pyogenes* **(Alamiri et al. 2020)**.

Although the gene H8F5G0 encodes for a hypothetical protein, the sequence is part of the gene scpA. This gene encodes for an important immune evasion protein C5-peptidase. ScpA (H8F5G0) is a multi-functional virulence factor with the potential of performing complement system-dependant and independent roles in *S. pyogenes* pathogenesis **(Lynskey et al. 2017)**. Furthermore, C5-peptidase and the two histidine triad proteins referenced above, HIT (A2RCZ2) and HIT (J7M9C9), all have the potential to reduce the effects of the human complement system via inhibiting the deposition of complement proteins in *S. pyogenes* **(Lynskey et al. 2017; Ogunniyi et al. 2009)**.

The M protein (JZM209) is a surface protein found on Group A *S. pyogenes* that has been shown to inhibit phagocytosis, interact with multiple host proteins, and exhibit antigenic diversity **(Smeesters et al. 2010)**. Not only is the M protein (JZM209) an antiphagocytic, but also has been shown to influence the coagulation cascade **(Paahlman et al. 2007)**. The M protein (JZM209) in *S. pyogenes* has been shown to increase pro-coagulant activity mediated by an upregulation of tissue factor on monocytes **(Paahlman et al. 2007)**. Furthermore, the study showed that *S. pyogenes* has developed a mechanism that evokes coagulation dysfunction that may not be limited to the site of the infection and could result in a systemic activation of the coagulation system **(Paahlman et al. 2007)**. This effect is possibly achieved by the ability of M protein (JZM209) to bind fibrinogen in plasma and mediate fibrinogen- and IgG-dependent platelet activation and aggregation, resulting in the release of granule proteins, upregulation of CD62P to the platelet surface, and complex formation with neutrophils and monocytes **(Palm et al. 2022)**.

Putative deoxyribonuclease (A2RDE7), also known by the gene name spd3 and protein names DNase C, Streptodornase C, and MF3 is a phage-encoded virulence factor of *S. pyogenes*. DNase (A2RDE7) has been shown to play a role in the destruction of extracellular traps produced by immune cells, such as neutrophils (NETs). Furthermore, Dnase (A2RDE7) has been shown to reduce TLR-9 signalling to dampen the immune response and produce cytotoxic deoxyadenosine to limit phagocytosis **(Remmington et al. 2018)**.

Finally, Lmb (Q9ZHG8) plays a crucial role in *S. pyogenes* adhesion to Epithelial cells **(Terao, Kawabata, Kunitomo, Nakagawa, et al. 2002)**. Zinc (Zn) is an essential trace element in living organisms and plays a vital role in the regulation of both microbial virulences and host immune responses **(Xia et al. 2021)**. Competition for zinc ions (Zn^{2+}) between the microbes and the hosts exists and Lmb (Q9ZHG8), HIT (A2RCZ2), and HIT (J7M9C9) have been shown to function as Zn scavengers for *S. pyogenes*, reducing the function of the human immune system by Zn deprivation **(Xia et al. 2021; Bellotti et al. 2021)**

Table 4.2: The conditional dependencies to Thrombomodulin and their significance in the GRaFO model prior to the pruning step via the application of the threshold in equation 4.14.

Clinical parameters	p-value	Human Genes	Associated Protein/function	p-value	Bacterial Genes	Associated Protein/function	p-value	Plasma analytes	p-value
Highest creatinine Day 1	***	SSH2	Slingshot Protein Phosphatase 2	***	Sp-H8F5G0	Hypothetical protein	***	IL-17A	***
Lactate BL	***	ABCA1	ATP Binding Cassette Subfamily A Member 1	***	Sp-A0A030GBR0	Hypothetical protein	**	ICAM-1	***
SOFA score Day 1	***	WDFY3	WD Repeat And FYVE Domain Containing 3	***	Sp-V6VJS2	30S ribosomal protein S19 domain protein	***	IL-18	***
SOFA renal Day 1	***	ADAMTS9	ADAM Metallopeptidase With Thrombospondin Type 1 Motif 9	**	Sp-C0M8M3	50S ribosomal protein L16	***	MMP-8	**
BMI	***	NCKAP1	NCK Associated Protein 1	***	Sp-A2RCZ2	Putative cell-cycle regulation histidine triad (HIT) protein	**	MPO	**
Standard Base Excess BL	***	DNAJC10	DnaJ Heat Shock Protein Family (Hsp40) Member C10	***	Sp-F5ZKC4	Ribosome-binding ATPase YchF	**	Pentraxin-3	***
Highest carbamide (1st 24hrs of ICU)	***	LYZ	Lysozyme	**	Sp-Q3K3U8	Adenylate kinase	***	Galectin-3	**
SAPS age	***	MTATP6P1	Mitochondrially Encoded ATP Synthase 6 Pseudogene 1	***	Sp-U9WVJ8	DnaK domain protein	*	TNFα	NS
Lowest pH Day 1	**	LAMA4	Laminin Subunit Alpha 4	***	Sp-U2J1P6	Uncharacterized protein	***	G-CSF	*
Lowest platelets Day 3	**	PDGFRA	Platelet Derived Growth Factor Receptor Alpha	***	Sp-Q3K3W0	30S ribosomal protein S17	***	I-α-1/COL1A1	*
pha BL	***	SRRM2	Serine/Arginine Repetitive Matrix 2	***	Sp-B5XIV8	Preprotein translocase subunit SecE	**	S100A8	NS
Subcutis dissolved - Surgical procedure 3	**	VPS13D	Vacuolar Protein Sorting 13 Homolog D	***	Sp-H8HBV4	Immunity factor for SPN	**	IL-4	*
Lowest pH Day 2	**	DOCK4	Dedicator Of Cytokinesis 4	**	Sp-K4Q8H6	Putative response regulator	***	Collagen-IVα1	NS
Platelet BL	**	TFRC	Transferrin Receptor	**	Sp-J7M9C9	Histidine triad protein	**	CCL-5/RANTES	NS
Age	**	FHL1	Four And A Half LIM Domains 1	***	Sp-Q9ZHG8	Metal ABC transporter substrate-binding lipoprotein/laminin-binding adhesin Lmb	***	IL-6	NS
Affection of upper arm at arrival (Specialised Hospital)	***	PEAK1	Pseudopodium Enriched Atypical Kinase 1	**	Sp-Q3K3X0	30S ribosomal protein S10	***	IL-1α	NS
Lowest bicoarbonate (1st 24hrs of ICU)	**	FNDC3B	Fibronectin Type III Domain Containing 3B	***	Sp-A2RDE7	Putative deoxyribonuclease	*	IL-22	NS
eGFR	**	SYNE1	Spectrin Repeat Containing Nuclear Envelope Protein 1	*	Sp-J7M209	M protein	**	CXCL-10/IP-10	NS
Body surface area	**	CFLAR	CASP8 And FADD Like Apoptosis Regulator	*	Sp-Q3K2F4	Peptide chain release factor 2	***	IL-13	NS
Adrenaline (highest dose) Day 1	*	HNRNPH1	Heterogeneous Nuclear Ribonucleoprotein H1	***	Sp-J7M6I8	Nucleoside permease	**	MMP-9	NS
Highest WBC Day 2	***	LAPTM5	Lysosomal Protein Transmembrane 5	**	Ec-R9VPP3	Hypothetical protein	**	S100A9	NS
Urine Output Day 3	*	SMG1	SMG1 Nonsense Mediated MRNA Decay Associated PI3K Related Kinase	***	Sp-Q48R07	Uncharacterized protein	**	CXCL-8/IL-8	NS
SOFA coag 1	***	IL1RN	Interleukin 1 Receptor Antagonist	NS	Sp-X5KJ99	Uridine phosphorylase	**	IL-1β	NS
Highest kalium (1st 24hrs of ICU)	**	CR1	Complement C3b/C4b Receptor 1 (Knops Blood Group)	*	Sp-J7M9C1	Trans-acting positive regulator	**	IL-2	NS
C-Reactive protein (highest) Day 1	*	SORL1	Sortilin Related Receptor 1	NS	Sp-Q5XAD4	Heat-inducible transcription repressor HrcA	***	IL-12p70	NS

Table 4.2: The conditional dependencies with clinical parameters, human genes, bacterial genes, and plasma analytes are separated into sub-tables. parameters in each sub-table are organised vertically based on the descending order of the associated mean decrease in gini and their significance in terms of p-values is given. There is no organisation or relationship present between parameters horizontally as the table is separated into sub-tables based on the paramater's provenance data set. The bacterial genes are UniRef90 Ids with a prefix of Sp or Ec referring to the gene belonging to *S. pyogenes* or *E. coli* respectively. The table can be visualised as a network with thrombomodulin as the parameter in the centre and the parameters in the table having conditional dependencies with thromobomodulin with the corresponding significance. The parameters in the table could be assumed to have secondary conditional dependencies with each other through thrombomodulin. Asterisks represent the significance of the conditional dependence with *** representing a p-value **< 0.01**, ** representing a p-value between **0.01** and **0.03**, * representing a p-value between **0.03** and **0.05**, and **NS** representing a p-value **> 0.05**

Mono- and poly-microbial NSTI

NSTI infections are often classified into mono-microbial or poly-microbial infections based on microbial aetiology**(D. L. Stevens and Bryant 2017)**. Mono-microbial infections (known as Type II) are caused by one causal micro-organism and poly-microbial infections (known as Type I) are caused by multiple causal organisms. Microbial aetiology was not a significant hub in the mixed graphical model, however, due to prior interest in the differences between patients, we look at the associations with microbial aetiology prior to applying the thresholds in the GraFo model in Table 4.3. The clinical parameters associated with microbial aetiology are related to the location of the infections and the severity of the infection. The parameter specimen sample site contains information regarding the location of the infection. The locations were divided into the categories of head & neck, upper arm, lower arm, hand, finger, thorax, abdomen, ano-genital area, upper leg, lower leg, foot, and toe. The plasma analytes were in general less significant, however, of the significant associations, IL-10, IL-22 & CXCL-10/IP-10 were identified as discriminatory plasma biomarkers for distinguishing the types of NSTI**(Medina et al. 2021)**. The GO enrichment analysis on the human genes enriches the Go terms positive regulation of plasma membrane repair, and negative regulation of hematopoietic stem cell differentiation in biological process and the GO terms palmitoyl-CoA 9-desaturase activity and S100 protein binding in molecular function. The pathway enrichment analysis enriches protein repair, activation of gene expression by SREBF (SREBP), NR1H2 and NR1H3 mediated signalling, inactivation of APC/C via direct inhibition of APC/C complex, autodegradation of Cdh1 by Cdh1:APC/C, and stimulation of the cell death response by PAK-2p34. The full results can be found in supplementary file S1 Chapter 4. The significant bacterial genes associated with microbial aetiology surprisingly all belong to *S. pyogenes*. Similar to LacA1 (Q99YH1), LacB1 (Q99YH2) is also involved in the tagatose-6 pathway for lactose and galactose metabolism that affects the EMP pathway downstream **(Pancholi et al. 2022)**. Glutamyl-tRNA synthase (P68890) is an essential gene recognised in the PATRIC database.

Table 4.3: The table shows the conditional dependencies to microbial aetiology and their significance in the GRaFO model prior to the pruning step via the application of the threshold in equation 4.14. the parameter microbial aetiology is a binary parameter that differentiates between mono-microbial (Type II) NSTI and poly-microbial (Type I) NSTI. The conditional dependencies with clinical parameters, human genes, bacterial genes, and plasma analytes are separated into sub-tables. Parameters in each sub-table are organised vertically based on the descending order of the associated mean decrease in gini and their significance in terms of p-values is given. There is no organisation or relationship present between parameters horizontally as the table is separated into four sub-tables based on the parameter's provenance data set. The bacterial genes are UniRef90 Ids with a prefix of Sp referring to the gene belonging to *S. pyogenes*. The table can be visualised as a network with microbial aetiology as the parameter in the centre and the parameters in the table having conditional dependencies with microbial aetiology with the corresponding significance. The parameters in the table could be assumed to have secondary conditional dependencies with each other through the parameter microbial aetiology. Asterisks represent the significance of the conditional dependence with *** representing a p-value < **0.01**, ** representing a p-value between **0.01** and **0.03**, * representing a p-value between **0.03** and **0.05**, and **NS** representing a p-value > **0.05**

Clinical parameters	p-value	Human Genes	Associated Protein/function	p-value	Bacterial Genes	Associated Protein/function	p-value	Plasma analytes	p-value
C-Reactive protein (highest) preoperative	***	MATR3	Matrin 3	***	Sp-Q99YH2	Galactose-6-phosphate isomerase subunit LacB 1	***	VCAM-1	*
Specimen sample site (Sample 1)	**	N4BP2L2	NEDD4 Binding Protein 2 Like 2	***	Sp-Q8K5R7	Uncharacterized protein	***	IL-22	**
Glucose BL	***	AHNAK	AHNAK Nucleoprotein	***	Sp-C5WEC0	DUF3013 family protein	***	Galectin-3	*
Days on mechanical ventilation (index ICU admission)	***	TXNIP	Thioredoxin Interacting Protein	**	Sp-C0M721	ATP synthase epsilon chain	***	TNFα	NS
Fascia - greyish discolouration at surgical procedure 1	***	SCD	Stearoyl-CoA Desaturase	**	Sp-B5XME5	Putative chorismate mutase	***	CXCL-10/IP-10	*
Affection of ano-genital region at arrival (Specialised hospital)	***	SAMD9L	Sterile Alpha Motif Domain Containing 9 Like	***	Sp-A2RFC8	ATP synthase subunit c	***	Fas-Ligand	NS
Highest systolic blood pressure (1st 24hrs in the ICU)	**	GDI2	GDP Dissociation Inhibitor 2	***	Sp-E4L6S2	Protein-export membrane protein SecG	***	IL-10	**
Frank Pus at surgical procedure 3	**	RIF1	Replication Timing Regulatory Factor 1	***	Sp-Q3JYQ4	Deoxyribose-phosphate aldolase	***	IL-36β/IL-1F8	NS
SAPS age	**	ZNF354B	Zinc Finger Protein 354B	**	Sp-B5XKQ3	DUF1146 domain-containing protein	***	CXCL-8/IL-8	NS
hgb BL	**	SORBS1	Sorbin And SH3 Domain Containing 1	**	Sp-P68890	Glutamyl-tRNA(Gln) amidotransferase subunit C	***	IL-33	NS

Mortality

NSTI infections have a relatively high mortality rate, ranging from 4% to 60% based on the time taken for diagnosis and surgery **(Nawijn et al. 2020)**. Looking at the associations with mortality in table 4.4, we find clinical variables have more significant conditional dependencies. Some of the associated clinical variables match the manually curated variables that were chosen to build a decision support system by Katz et al. **(S. Katz et al. 2022)**. Conversely, the conditional dependencies with plasma analytes are not significant. This may suggest that biological variables and measurements may not hold the ultimate predictive power for mortality predictions.

GO enrichment analysis on the human genes reveals the GO terms acetylcholine-mediated vasodilation involved in regulation of systemic arterial blood pressure and basophil chemotaxis in biological process and the GO terms α- 1,3- mannosylglycoprotein 2- β- N- acetylglucosaminyltransferase activity, beta-3 adrenergic receptor

activity, and insulin-like growth factor receptor activity in molecular functions. The pathway enrichment analysis reveals cytosolic iron-sulphur cluster assembly, ion influx/efflux at host-pathogen interface, SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion, neurophilin interactions with VEGF & VEGFR, VEGF binds to VEGFR leading to receptor dimerisation, expression and processing of neutrophins, cellulat response to hypoxia, regulation of BCH1 activity, NFE2L2 regulates pentose phosphate pathway, defective homologous recombination repair (HRR) due to PALB2 & BRCA1 loss of function, and impaired BRCA2 binding to PALB2 as enriched pathways.

Conditional dependencies to bacterial genes contain genes from *S. pyogenes* and *S. dysgalactiae*. When we study the bacterial genes, we find genes representing virulence factors, drug targets, essential genes and transporter proteins. PTS family ported component HPr (F8IR47) is involved in the interaction with CcpA (Q8E0M3) in the negative regulation of Streptococcal surface dehydrogenase (SDH) and has been recognised as a virulence factor and a drug target **(Jin et al. 2011)**. Furthermore, it is an essential co-factor for Ccpa-regulation **(DebRoy et al. 2021)**. Streptolysin (C5WFP7) a transport protein is also involved in the regulation of SDH. The gene also named sagG forms part of the Streptolysin S-operon in *S. pyogenes*, an eight-gene operon that produces the potent toxin Streptolysin S **(E. M. Molloy et al. 2011)**. Streptolysin S (C5WFP7) has been known to contribute to the intracellular iron acquisition by *S. pyogenes* from the host via the process of hemolysis of the host red blood cells **(E. M. Molloy et al. 2011)**. Furthermore, Streptolysin S (C5WFP7) has been shown to contribute to the diminishing of the host's immune response by disrupting the ability of the host cell to produce signals that are chemo-tactic for neutrophils **(E. M. Molloy et al. 2011)**.

Sic1.01 (K4N5N3) is also known as Streptococcal inhibitor of complement (SIC). It is a secreted protein that has been found to interact with CD14 and TLR2 on monocytes, triggering the activation of NF-$\kappa\beta$ and p38 MAPK pathways with the release of TNFα and IFN-γ **(Neumann et al. 2021)**. Streptopain (P0C0J1) also known as streptococcal pyogenes exotoxin B (SPE B) is a cysteine protease that plays a crucial role in the maturation of pro-SPE B zygomen **(C.-Y. Chen et al. 2003)**. CcpA (Q8E0M3) and Hpr (F8IR47) have been shown to be involved in the process of sensing and adapting to a dynamic host environment.

CcpA (Q8E0M3) and Hpr (F8IR47) have been shown to regulate the expressions of the proteins SpeB (P0C0J1), Sic (K4N5N3), Streptolysin S (C5WFP7), NAD^+-glycohydrolase (H8HBV4), and Lmb (Q9ZHG8) **(Loughman et al. 2006)**. As discussed in regards to HrcA (Q5XAD4) in section *Thrombomodulin and differences in coagulation in NSTI patients*, CcpA (Q8E0M3) has also been implicated in the regulation of biofilm in various micro-organisms such as *Streptococcus gordonii* and *S. aureus* **(L. Zheng et al. 2012; Seidl et al. 2008)**. *S. pyogenes* biofilm formation has been positively linked with the expression of NAD^+-glucohydrolase (H8HBV4). Furthermore, the expressions of Spe B (P0C0J1), Streptolysin S (C5WFP7), and M protein (JZM209) was shown to be lowered in biofilm bacteria suggesting a potential role for these factors in the biofilm formation **(Alamiri et al. 2020)**. 50S ribosomal protein L7/L12 (Q48TS3) is recognised as an essential gene and Putative ATP-binding cassette transporter-like protein (G4R2T6) is recognised to be a transporter protein in the PATRIC database.

Table 4.4: The conditional dependencies to mortality and their significance in the GRaFO model prior to the pruning step via the application of the threshold in equation 4.14. the parameter mortality is a binary parameter that holds the information on whether or not the NSTI patient survived. The conditional dependencies with clinical parameters, human genes, bacterial genes, and plasma analytes are separated into sub-tables. Parameters in each sub-table are organised vertically based on the descending order of the associated mean decrease in gini and their significance in terms of p-values is given. There is no organisation or relationship present between parameters horizontally as the table is separated into four sub-tables based on the parameter's provenance data set. The bacterial genes are UniRef90 Ids with a prefix of Sp or Sd referring to the gene belonging to *S. pyogenes* or *S. dysgalactiae* respectively. The table can be visualised as a network with mortality as the parameter in the centre and the parameters in the table having conditional dependencies with microbial aetiology with the corresponding significance. The parameters in the table could be assumed to have secondary conditional dependencies with each other through the parameter microbial aetiology. Asterisks represent the significance of the conditional dependence with *** representing a p-value **< 0.01**, ** representing a p-value between **0.01** and **0.03**, * representing a p-value between **0.03** and **0.05**, and **NS** representing a p-value **> 0.05**

Clinical parameters	p-value	Human Genes	Associated Protein/function	p-value	Bacterial Genes	Associated Protein/function	p-value	Plasma analytes	p-value
Highest lactate Day 2	***	SLC11A1	Natural Resistance-Associated Macrophage Protein 1	**	Sd-Q3K3W0	30S ribosomal protein S17	**	I-α-1/COL1A1	*
Lowest Natrium (1st 24hrs in ICU)	***	PTP4A1	Protein Tyrosine Phosphatase 4A1	**	Sp-Q3K3W0	30S ribosomal protein S17	NS	CCL-2/MCP-1	**
Lowest Standard Base Excess Day 2	***	NRSN1	Neurensin 1	**	Sd-C5WFP7	Streptolysin S export ATP-binding protein	**	IL-33	NS
Urine output (24hrs) Day 2	***	FTH1P2	Ferritin Heavy Chain 1 Pseudogene 2	**	Sd-B5XLU2	DUF2273 domain-containing protein	**	Collagen-IVα1	NS
Albumine (24hrs) BL	***	WDR74	WD Repeat-Containing Protein 74	**	Sd-Q3K3U6	30S ribosomal protein S13	**	IL-17A	NS
Lowest pH Day 2	***	RPS27L	Ribosomal Protein S27 Like	NS	Sd-F8IR47	PTS family porter component HPr	**	CCL-5/RANTES	NS
L_Na	***	DOCK5	Dedicator Of Cytokinesis 5	*	Sp-P0C0J1	Streptopain	NS	C5/C5a	NS
Max infusion rate of Adrenaline Day 2	***	FOSL2	FOS Like 2, AP-1 Transcription Factor Subunit	**	Sp-Q8K763	Phage protein	**	IL-1R2	NS
Application of vaccum assisted closure: surgical procedure 3	**	PCSK5	Proprotein Convertase Subtilisin/Kexin Type 5	NS	Sd-U3TN16	Periplasmic component of efflux system	**	IL-22	NS
Age	***	VEGFA	Vascular Endothelial Growth Factor A	*	Sd-E7PVR9	Hypothetical ribosome-associated protein	**	IL-36β/IL-1F8	NS
Max infusion rate of Noradrenaline Day 3	***	MGAT1	N-Glycosyl-Oligosaccharide-Glycoprotein N-Acetylglucosaminyltr I	**	Sd-E7PUV1	PspC domain-containing protein	*	S100A9	NS
Highest Carbamide (1st 24hrs in ICU)	**	ZNF354B	Zinc Finger Protein 354B	*	Sd-Q48TS3	50S ribosomal protein L7/L12	*	Thrombomodulin	NS
SAPS II	***	SOD2	Superoxide Dismutase 2	**	Sd-C5WEM6	Murein hydrolase regulator	**	IL-18	NS
SOFA (missing replaced with max values) Day 3	**	IGF2R	Insulin Like Growth Factor 2 Receptor	NS	Sd-E8QCI7	Antiholin-like protein	**	IL-1R1	NS
Upper leg amputated	***	ZFP36L1	Zinc Finger Protein 36, C3H Type-Like 1	NS	Sd-P66202	50S ribosomal protein L31 type B	*	IL-6	NS
Colloids BL (including albumin) (24hrs)	***	LDHA	Lactate Dehydrogenase A	NS	Sp-K4N5N3	Sic1.01	*	CXCL-8/IL-8	NS
Blood Products BL (24hrs)	***	ARRDC3	Arrestin Domain Containing 3	NS	Sd-P66376	30S ribosomal protein S12	*	S100A8	NS
RBCs Day2	***	NBPF14	Neuroblastoma Breakpoint Family Member 14	NS	Sd-G4R2T6	Putative ATP-binding cassette transporter-like protein	*	ICAM-1	NS
Preadmission creatinine (from MDRD equation)	**	BACH1	BTB Domain And CNC Homolog 1	NS	Sd-P0DE26	50S ribosomal protein L24	*	Il-2	NS
RRT at baseline	**	LRRK1	Leucine Rich Repeat Kinase 1	NS	Sd-C0M7I0	50S ribosomal protein L27	*	IL-10	NS

4.3.4 Stratifying patients based on infection location and aetiology

An effort was undertaken to stratify patients according to the site of infection and microbial aetiologies. We employed a GGM model to try and uncover inter-bacterial gene associations in these patients. Nevertheless, the statistical significance of the results could not be determined due to the insufficient sample sizes in the stratified groups.

The GGM models were utilised to derive insights into the following groups

- Difference in the associations between *Fusobacterium nucleatum* and human genes between Type I and Type II infections.
- Inter-bacterial gene associations between *F. nucleatum*, *Streptococcus intermedius*, *Aggregatibacter aphrophilus*, *Streptococcus infantis*, *Streptococcus constellatus*, and *Streptococcus milleri*
- Associations between *S. infantis*, *S. milleri*, and *E. coli* in patients with pelvic abdominal infections.
- Associations between *Clostridium septicum*, *P. micra*, *Porphyromonas endodontalis*, *Porphyromonas gingivalis*, *Porphyromonas uenonis*, and *Porphyromonas asaccharolytica* in patients with head and neck infections.
- Difference in the associations between *C. septicum* and human genes between Type I and Type II infections.

In the first group, no *F. nucleatum* genes were expressed in Type II NSTI samples, whereas in Type I NSTI samples, PduA (A0A140PXV6), a known virulence factor was detected. Nonetheless, the sample size of poly-microbial patients expressing *F. nucleatum* genes was extremely limited. In the second group, neither *A. aphrophilus* nor *S. milleri* genes were expressed in any of the samples, and a limited number of samples expressed genes from the remaining bacteria at the same time. Analysis of inter-bacterial gene associations did not identify any virulence factors from *S. intermedius*, although PduA (A0A140PXV6) was once again found in *F. nucleatum*. Simultaneously, a high connectivity measure was observed for the Oxidoreductase transporter (B1IAL3) from *S. infantis*. However, the restricted sample size (n=7) rendered the interpretation of these results impossible.

No *S. intermedius* and *S. milleri* genes were expressed in patients afflicted with pelvic abdominal infections. Patients with head and neck infections did not express any *C. septicum* genes. Research on virulence factors for *P. endodotalis*, *P. uenonis*, and *P. asaccharolytica* remains relatively unexplored. β-galactosidase (B2RJL9), however, displayed high connectivity as a virulence factor in *P. gingivalis*. As before, the sample sizes were too low to provide statistically reliable interpretations. Lastly, no *C. septicum* genes were found in either Type I or Type II NSTI samples. Complete information on all the genes from the different groups can be found in supplementary file S?

4.3.5 Limitations of the study

This study is an exploratory affair to study associations between existing multiomics datasets and genrate hypothesis to explore underlying mechanisms in NSTI. Limita-

tions in this study are a result from both computational methods and data samples. Limitations from the samples include relatively small number of samples, heterogeneity among the patients (such as comorbidity), and time of admission to the hospital. One of the limitations with the mixed graphical model comes from ranking the results of mixed data types where when predicting for discrete variables, each level contributes and additive effect while predicting a continuous variable, each level contributes linear effects. Other limitation in the MGM is that it mirrors the limitations presented in linear and logistic regressions when dealing with multicollinearity and the lasso regularisation (λ) can correct for overfitting but the model can not be made more stringent by constantly increasing the λ value. Limitations in the GRaFo model include its inability to associate values and directionality to the conditional dependence calculated from the model.

However, clinicians using standardised SOPs, a similar prospective observational study design in all clinical centres, using multiple complementary MGM methods that counter each other's limitations in ensemble and a statistically stringent application of all the models reinforces the study design.

4.4 Conclusions

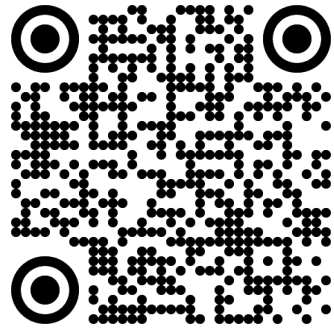
Through a systems biology approach, we endeavour to decipher the association structures formulated by clinical variables, bacterial and human genes, and plasma analytes measured in NSTI patients, thereby deepening our understanding of the comprehensive mechanisms and responses that influence phenotypic observations. Leveraging methodologies designed for mixed data types and limited sample sizes, we propose a series of data-driven hypotheses for subsequent experimental evaluation. The derived MGM network emphasises the interplay between *S. pyogenes* and neutrophils. Our findings highlight the association of mortality with the regulation of SpeB, Streptolysin S, and Sic by Catabolic control protein A, operating through the co-factor Hpr to control biofilm expression. Moreover, we observe Thrombomodulin to be associated with numerous *S. pyogenes* genes previously recognised for their interactions with the human immune system and blood constituents. The elucidation of key pathogenic mechanisms invariably contributes to the further development of interventional strategies and therapeutic approaches.

4.5 Author contributions

SJ and ES designed the study; SJ and YB performed the analysis; SJ, OO, KAM, HR, TB, MS, OH, ANT, VAdP, SS and ES provided advice and interpretation of results; SJ visualised the results; SJ wrote the manuscript; OO, KAT, MS, ANT, VAdP, SS and ES critical revision. All authors reviewed the manuscript.

4.6 Funding

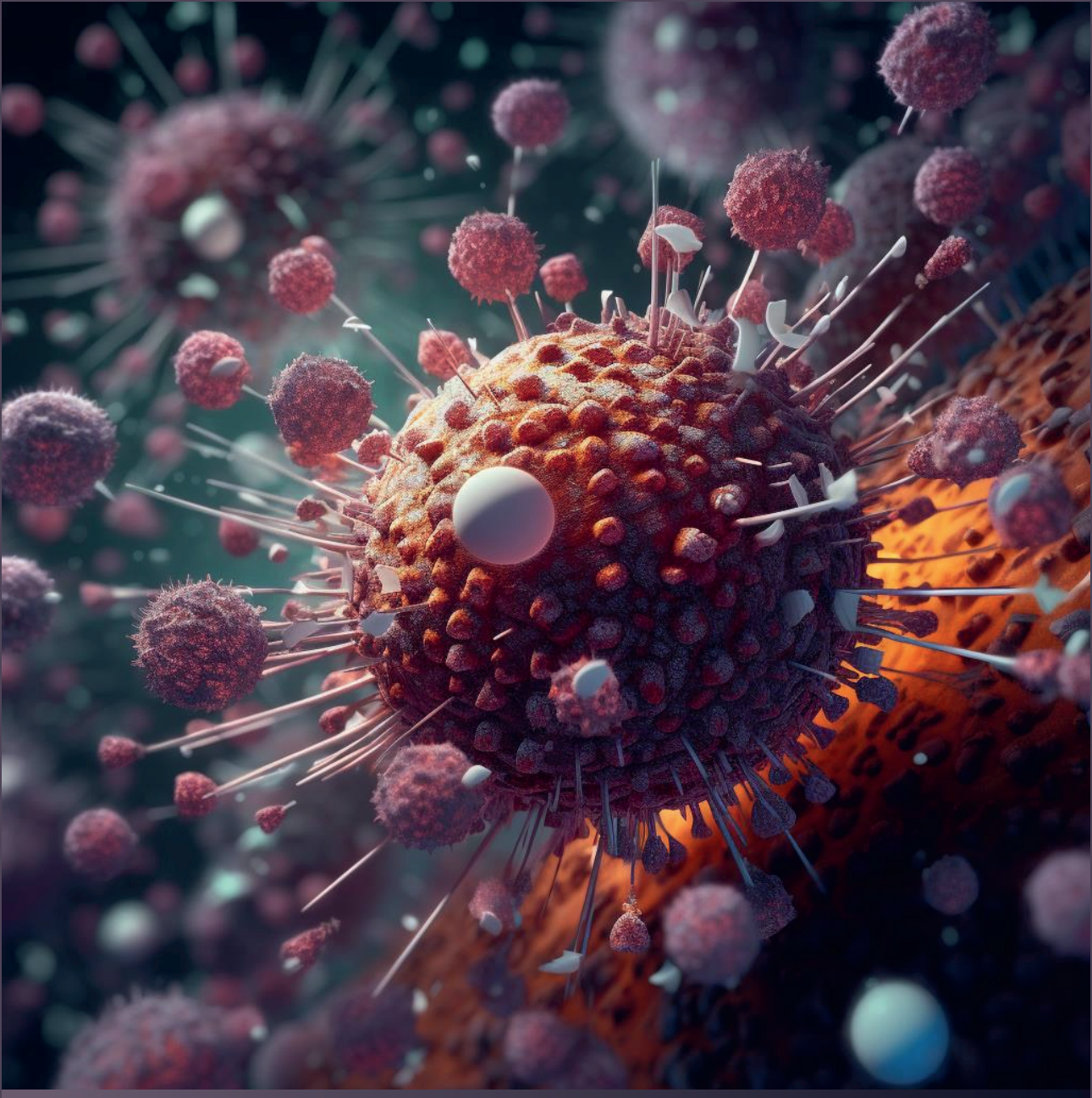
This work was supported by the Center for Innovative Medicine (CIMED) and Region Stockholm (no. 20180058); the Swedish Research Council (2018-02475); the Euro-



Publication

pean Union Seventh Framework Programme (FP7/2007-2013) under the grant agreement 305340 (INFECT project); the Swedish Governmental Agency for Innovation Systems (VINNOVA), Innovation Fund Denmark (8114-000005B) and the Research Council of Norway under the frame of NordForsk (project no. 90456, PerAID); the Swedish Research Council, Innovation Fund Denmark (8113-000009B), the Research Council of Norway, the Netherlands Organisation for Health Research and Development (ZonMW) and DLR Federal Ministry of Education and Research, under the frame of ERA PerMed (project 2018-151, PerMIT); and the Swedish Children's Cancer Foundation (TJ2018-0128).

Chapter 5 Chapter



Sanjeevan Jahagirdar[1], Lorna Morris[8], Nirupama Benis[1,9] Oddvar Oppegaard[2], Mattias Svenson[4], Ole Hyldegaard[5,6], Steinar Skrede[2,3], Anna Norrby-Teglund[4], INFECT Study group[#], Vitor A. P. Martins dos Santos[7,8], Edoardo Saccenti[1]

#INFECT study group (Trond Bruun, Eivind Rath, Torbjørn Nedrebø, Per Arnell, Anders Rosen, Morten Hedetoft, Martin B. Madsen, Mattias Svensson, Johanna Snäll, Ylva Karlsson, & Michael Nekludov)

Turn to page 377 for author affiliations

This chapter is adapted from:

Jahagirdar, S., Morris, L., Benis, N., Oppegaard, O., Svenson, M., Hyldegaard, O., Skrede, S., Norrby-Teglund, A., Bruun, T., Rath, E., Nedrebø, T., Arnell, P., Rosen, A., Hedetoft, M., Madsen, M. B., Svensson, M., Snäll, J., Karlsson, Y., Nekludov, M., … Saccenti, E. (2022). Analysis of host-pathogen gene association networks reveals patient-specific response to streptococcal and polymicrobial necrotising soft tissue infections. ***BMC Medicine*****, *20*(1), 173. https://doi.org/10.1186/s12916-022-02355-8**

Inside the Inferno: An Inspection of the Host-Pathogen Interactions

Abstract

Background: Necrotising soft tissue infections (NSTIs) are rapidly progressing bacterial infections usually caused by either several pathogens in unison (polymicrobial infections) or *S. pyogenes* (mono-microbial infection). These infections are rare and are associated with high mortality rates. However, the underlying pathogenic mechanisms in this heterogeneous group remain elusive.
Methods: In this study, we built interactomes at both the population and individual levels consisting of host-pathogen interactions inferred from dual RNA-Seq gene transcriptomic profiles of the biopsies from NSTI patients.
Results: NSTI type-specific responses in the host were uncovered. The *S. pyogenes* mono-microbial subnetwork was enriched with host genes annotated with involved in cytokine production and regulation of response to stress. The polymicrobial network consisted of several significant associations between different species (*S. pyogenes*, *P. asaccharolytica* and *E. coli*) and host genes. The host genes associated with *S. pyogenes* in this subnetwork were characterised by cellular response to cytokines. We further found several virulence factors including hyaluronan synthase, Sic1, Isp, SagF, SagG, ScfAB-operon, Fba and genes upstream and downstream of EndoS along with bacterial housekeeping genes interacting with the human stress and immune response in various subnetworks between host and pathogen.
Conclusions: At the population level, we found aetiology-dependent responses showing the potential modes of entry and immune evasion strategies employed by *S. pyogenes*, congruent with general cellular processes such as differentiation and proliferation. After stratifying the patients based on the subject-specific networks to study the patient-specific response, we observed different patient groups with different collagens, cytoskeleton and actin monomers in association with virulence factors, immunogenic proteins and housekeeping genes which we utilised to postulate differing modes of entry and immune evasion for different bacteria in relationship to the patient's phenotype.

5.1 Background

Necrotising soft tissue infections (NSTI) are devastating bacterial infections characterised by impairment and injury in any layer of the soft tissue compartment, extending from the epidermis to the deep musculature **(Goldstein et al. 2007; Bonne et al. 2017)**. These infections are relatively rare (0.2 to 15.5 per 100,000 people/year) but their aggressive nature poses severe threats due to the high risk of mortality and long-term disability which often results from the extensive tissue loss and amputations often prescribed to control the infections **(D. L. Stevens and Bryant 2017)** due to the fact that progression is rapid, and early diagnosis is vital for improving the prognosis of affected patients **(Pham et al. 2009; A. K. May 2011; M. B. Madsen, Skrede, et al. 2019; T. Chan et al. 2008)**. NSTIs can be caused by either a single bacterial species (monomicrobial NSTI, or Type II NSTI) or by multiple species (polymicrobial NSTI, or Type I NSTI), and the relative occurrence of the two types of NSTI differs significantly based on the geography and patient characteristics **(D. L. Stevens and Bryant 2017; Skrede et al. 2020)**.

S. pyogenes is the most common pathogen in monomicrobial NSTIs **(D. L. Stevens and Bryant 2017)**, but other streptococcal species **(Bruun, Kittang, et al. 2013)** and *S. aureus* are also known to cause monomicrobial NSTIs **(Miller et al. 2005)**. In this study, we only focus on type II NSTI caused by *S. pyogenes*, as well as polymicrobial NSTIs that are associated with a mixture of obligate anaerobic and facultative anaerobic bacteria **(Cocanour et al. 2017)**, such as *Enterobacteriaceae*, *Bacteroides* spp., *Porphyromonas* spp., *Prevotella* spp., *Peptostreptococcus* spp. and *Clostridium* spp. **(Goldstein et al. 2007; Elliott et al. 2000)**.

Monomicrobial NSTIs caused by *S. pyogenes* have been studied extensively and many of the virulence factors and toxins expressed by the bacterium to colonise the host tissue and bypass the host immune defences have been characterised **(Johansson, Thulin, et al. 2010)**. In contrast, the pathogenic strategies and the complex dynamics of bacterial communities underlying polymicrobial NSTIs are poorly understood. One of the major limitations in our understanding of NSTI at the molecular level (and of bacterial infections in general) is the insufficient information about the web of molecular interaction, also known as interactome, between pathogens and the human host. In contrast with Mendelian diseases, where one or few genes can be directly linked to the disease, bacterial infections arise from the complex interactions between bacteria, the host immune system, predisposition, risk and environmental factors **(Doron et al. 2008)**. Interactomics focuses on the representation and the analysis of the interactions between biological features on a global scale **(T. Ito et al. 2001)** using network approaches to simplify a complex system like, in this case, a bacteria-host system, and to summarise it as components (nodes) and interactions (edges) between them **(Vidal et al. 2011)**. Both nodes and edges can be different in nature, depending on the type of interactome considered. In this study, nodes are human and bacterial genes and edges represent the existence of a correlation between the expression profiles of these genes, thus representing the mutual response of host and pathogen and providing a global view of the observed interactions at the molecular level **(Alonso-Lopez et al. 2016)**.

The present study builds and expands (on) the data obtained in the largest cohort

study of NSTI patients in the world to date, the INFECT study. Thänert et al. **(Thänert et al. 2019)** analysed dual RNA sequencing of NSTI patient biopsies together with microbial community profiling using 16S rRNA sequencing data **(Thänert et al. 2019)** collected within the INFECT study **(M. Madsen et al. 2018)**, and showed that gene expression profiles of tissues from NSTI patients differed significantly between monomicrobial streptococcal and polymicrobial infections, identifying the core inflammatory signatures in both instances.

Here, we used network analysis to explore relationships between co-expressed host and bacterial gene pairs, complementing and expanding the results of Thänert et al. **(Thänert et al. 2019)** by considering a larger number of samples with the aim to provide insight into the interaction and dynamics between the pathogens and the host and illuminate some of the underlying molecular mechanisms in the pathophysiology of NSTI using a systems biology approach **(T. Ito et al. 2001; Alonso-Lopez et al. 2016)**.

5.2 Methods

5.2.1 Study design

The study is founded in the INFECT study on clinical and pathogenesis in NSTI, where patients were included by prospective enrolment through 4.5 years in five Scandinavian referral hospitals (see table 5.1). Study design and presentation of clinical results are detailed elsewhere **(M. B. Madsen, Skrede, et al. 2019; M. Madsen et al. 2018; Bruun, Rath, et al. 2021)**. The INFECT study is registered at ClinicalTrials.gov (NCT01790698).

Tissue biopsies and plasma samples on the day of hospital admission (day 0) were obtained from patients diagnosed with NSTI and admitted to Karolinska University Hospital in Stockholm, Copenhagen University Hospital, Blekinge County Council Hospital in Karlskrona, Sahlgrenska University Hospital in Gothenburg and Haukeland University Hospital in Bergen in the framework of the EU project INFECT (https://permedinfect.com/projects/peraid/).

Diagnosis of NSTI was based on the presence of necrotic or deliquescent soft tissue with widespread undermining of the surrounding tissue. Patients were excluded in the absence of reports of necrotic or deliquescent tissue. More details on patient characteristics and study design can be obtained from **(M. Madsen et al. 2018)**.

5.2.2 Ethical considerations

The INFECT study was conducted in accordance with the Declaration of Helsinki and was approved by the regional Ethical Review Board at the Karolinska Institutet in Stockholm, Sweden (Ethics Permits: 2012/2110-31/2), the National Committee on Health Research Ethics in Copenhagen, Denmark (Ethics permits: 1151739), the regional Ethical Review Board in Gothenburg, Sweden (Ethics permits: 930-12) and Bergen, Norway (2012/2227/REC West). All experiments were performed in accordance with the approved ethics applications specified above. All patients provided written informed consent.

Table 5.1: Clinical parameters associated with the patients whose biopsies were analysed with dual RNA-seq and used to build the host-pathogen interactome in NSTI. Median and Interquartile ranges (Lower Quartile – Upper Quartile) are given. Under the section of "hospital" we show both the number of patients admitted and the number of biopsies taken (in brackets) from the hospitals. The number of patients may not coincide with the number of biopsies since multiple biopsies can be taken from the same patient. For instance, "13 (15)/35 (42)" means that 13 out of the 35 patients and 15 out of the 42 biopsies were from Righospitalet. Since patients can have infections in multiple locations, the percentage may not add up to 100%.

	Monomicrobial NSTI	**Polymicrobial NSTI**
Age (years)	57 (44 - 61)	55 (46.5 - 64)
Sex		
Female/Male (%)	12 (34.3%)/ 23 (65.7%)	9 (39.2%)/ 14 (60.8%)
Outcome		
Mortality 30 days (%)	1 (2.86%)	4 (17.4%)
Mortality 90 days (%)	3 (8.57%)	4 (17.4%)
Mortality 365 days (%)	6 (17.14%)	6 (26.1%)
Hospital		
Righospitalet Copenhagen	13 (15)/35 (42)	11 (11)/ 24 (25)
Karolinska University Hospital	7 (8)/35 (42)	8 (8)/ 24 (25)
Sahlgrenska University Hospital	5 (6)/35 (42)	2 (2)/ 24 (25)
University of Bergen	10 (13) /35 (42)	2 (3)/ 24 (25)
Laboratory values		
Haemoglobin ($\frac{g}{dl}$)	10.5 (2.51 - 11.76)	8.8 (7.9 - 10.55)
White blood cells ($\frac{10^9}{l}$)	13.4 (11.15 - 16.625)	12.3 (8.7 - 15.7)
C-reactive protein ($\frac{mg}{l}$)	207 (152 - 293.25)	293 (140 - 343)
Creatinine ($\frac{\mu mol}{l}$)	123 (87.75 - 233.75)	116.5 (81.25 - 180.75)
SOFA score	9 (5 - 11)	8 (6 - 11)
SAPS II	40 (35 - 51)	40.5 (29 - 48)
Location of Infection in patients		
Head & neck (%)	7 (20%)	9 (39.1%)
Upper extremities including thoracic involvement (%)	14 (40%)	3 (13%)
Abdomen & ano-genital area (%)	5 (14.3%)	12 (52.2%)
Lower extremities (%)	15 (42.9%)	5 (21.7%)

5.2.3 Experimental methods

RNA-seq sample preparation and sequencing

This study makes use of RNAseq samples used in **(M. Madsen et al. 2018; Thänert et al. 2019)** plus additional data not available at the time. All samples were handled and processed as described in **(M. Madsen et al. 2018; Thänert et al. 2019)**.

Sample selection

In line with the study by Thänert et al. **(Thänert et al. 2019)**, we retained all those subjects/samples for which dual RNA-seq data (i.e. transcriptomics data for both host and pathogen) was available together with 16S bacterial rRNA gene sequencing data collected on the day of hospital admission: this results in 81 samples available for analysis (Figure 5.1).

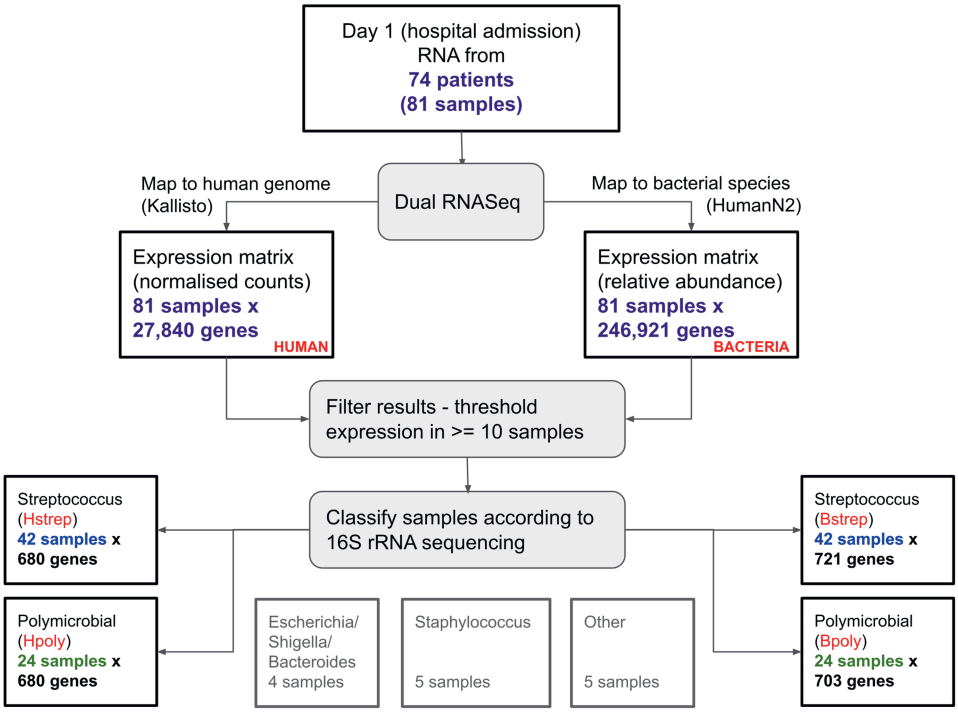


Figure 5.1: Flowchart for sample selection, mapping and filtering of dual RNA-seq for the generation of gene expression matrices for *S. pyogenes* monomicrobial infections (Hstrep - Human, Bstrep -Bacteria) and polymicrobial infections (Hpoly – Human, Bpoly -Bacteria). Bacterial and Human gene expression are measured on the same tissue biopsies from NSTI patients. Clinical information associated with patients and samples\biopsies is given in Table 1

We followed the classification established by Thanert et al.**(Thänert et al. 2019)** who assigned the 81 samples to 5 different types of infection based on their associated bacterial composition according to 16S rRNA gene sequencing, namely Staphylococcus ($n = 5$), Streptococcus ($n = 42$), Polymicrobial ($n = 25$), Escherichia/Shigella/Bacteroides ($n = 4$) and other ($n = 5$). Classification was based on average-linkage hierarchical agglomerative clustering using the relative abundance of the identified bacterial communities. The optimal number of clusters in the resulting sample dendrogram was determined using the J-index and distinct specimen clusters were defined to represent different types of NSTIs. This information was obtained from the Supplementary Data 3 from **(Thänert et al. 2019)**. The sample size for *Staphylococcus*, *Escherichia/Shigella/Bacteroides* and others was not large enough to build robust correlation networks, thus we focused on the two larger groups, namely samples from patients with *S. pyogenes* and polymicrobial NSTI. We indicate with Hstrep and Hpoly the matrix of human gene expression measured on biopsies obtained from patients with diagnosis streptococcal NSTI and polymicrobial NSTI, respectively; we indicate

with Bstrep and Bpoly the matrix of bacterial gene expression for streptococcal NSTI and polymicrobial NSTI, respectively.

5.2.4 Bioinformatics analysis

RNA sequences were mapped against both human and bacterial genomes to obtain gene expression of both the host and resident pathogens. Quality control was performed with the tool FASTQC **(Andrews et al. 2010)**. The mapping tool Kallisto4 was used to map the sequences against the human genome (GRCh38 release 91). The resulting read counts, from TSV files, were loaded into R using the package tximport **(Soneson et al. 2015)**.

The same sequences were also mapped against several bacterial genomes using the published pipeline HUMAnN2 (version 0.11.1) **(Franzosa, L. J. McIver, et al. 2018)** which uses a series of tools to map RNA sequences to bacterial proteins from the database UniRef **(Suzek et al. 2007)**. HUMAnN2 uses a tiered search for taxonomic profiling. Firstly, it searches for a pre-selected set of marker genes (from the MetaPhIAn2 database) unique for each species. Secondly, it maps all the remaining reads to the pangenomes of the detected species. Thirdly, the remaining unclassified reads are mapped to a protein sequence database (DIAMOND). We used the default coverage threshold in HUMAnN2 in order for a sequence in a particular species to be detected. Alignments are not considered if they do not pass the default coverage threshold. We opted to use the recommended database UniRef90. Uniref90 identifiers have the form UniRef90_ID, where ID is a UniProtKB accession or UniParc identifier of the representative sequence from the cluster of 90% identical sequences referred to by UniRef90_ID. For ease of discussion and visualisation in this manuscript, we refer to the ID part only and use functional annotation from Uniprot where available to describe these sequences.

The mapped human and bacterial data were post-processed and filtered before integration with each other. Filters were set for the level of gene expression and this threshold had to be met in at least 10 samples to focus on gene expression patterns in host and pathogen that are predominant in several patients. Since not all genes are expressed or can be in all samples, we needed to ensure that a gene was expressed in a sufficiently large number of samples to avoid spurious associations. We found that 10 samples were sufficient to ensure robustness of the results. In addition, genes mapped from *E. coli* were also removed from samples classified as 'Streptococcus' according to 16S sequencing. The final data sets for the Streptococcus monomicrobial classified samples contained 680 human genes and 721 bacterial Uniref90 sequences, and the poly-microbial classified samples contained 680 human genes and 703 bacterial Uniref90 sequences. The gene-level abundance was calculated as reads per kilobase units (RPKU), as defined by HUMAnN2.

5.2.5 Data transformation

Data was transformed to stabilise the variance of the data. Gene expression values were transformed taking the square root of the original values.

5.2.6 Predictive modelling

The Random Forest (RF) algorithm was used to classify monomicrobial and polymicrobial patients on the basis of human gene expression profiles **(Breiman 2001)**. To reduce the potential bias due to an unbalanced number of subjects/samples per group, we imposed a number of k=100 resampling, considering the 85% of data to retain for each compared group. Accuracy, sensitivity, specificity, and related 95% CI of all performed models were assessed according to the standard definitions and were determined by means of permutated test (k=1000 times). For all calculations, the "randomForest" function, implemented in the R package Random Forest, was used to grow a decision forest composed of 1000 trees. Default parameters were used.

5.2.7 Gene-gene association network inference

Association networks between human and bacterial genes were built using the Probabilistic Context Likelihood of Relatedness on Correlation (PCLRC) algorithm **(Saccenti, Suarez-Diez, et al. 2015)** which is based on the original CLR algorithm **(Faith et al. 2007)** and has been shown to be robust against variation in sample size and noise **(Jahagirdar, Suarez-Diez, et al. 2019)**. In the current study, we replaced standard correlations between two molecular features i and j (either metabolites or genes) with partial correlations obtained using a Gaussian Graphical Model (GMM). The PCLRC algorithm uses resampling to estimate robust correlation based on the Context Likelihood of Relatedness approach which estimates the relevance of the associations between two features by considering background associations. The PCLRC returns a probability matrix P, containing the likelihood $0 < p_{ij} < 1$ of each observed association r_{ij} between each gene pair.

All p-values were corrected for multiple testing using the Benjamini & Hochberg method **(Benjamini et al. 1995)**. Only corrected P-values (P_adjust) smaller than 0.05 were retained in the analysis to give us a partial correlation network of genes filtered for only the most significant associations. Significant associations r_{ij} between the i^{th} host and j^{th} pathogen gene were defined in equation 5.1. Default parameters were used (number of resampling iterations Niter = 1000; the fraction of the samples to be considered at each iteration fraction = 0.75 and fraction of the total predicted interactions to be kept at each iteration rank.thr = 0.3). All networks are undirected and represented as an $m \times m$ adjacency matrix M, populated by association (edges) between genes i and j (nodes). For each node in the $m \times m$ network(s) we calculated the node degree (connectivity) as shown in equation 5.2.

5.2.8 Estimation of partial correlations using Gaussian graphical model

Partial correlations were estimated using a Gaussian Graphical Model (GGM) as implemented in the GeneNet R package **(Opgen-Rhein et al. 2007; J. Schäfer, Opgen-Rhein, et al. 2006)**. GeneNet allows estimating a GGM from a small sample of high-dimensional data in a computationally and statistically efficient way. It uses an analytic shrinkage estimation of covariance and partial correlation matrices and performs optimal model selection based on local false discovery rate multiple testing. The edges (i.e., the associations) to be included in the final association network are

selected using a computational algorithm depending on the relative values of the pairwise partial correlations. For more details on GeneNet implementation, we refer to the original publication **(J. Schäfer, Opgen-Rhein, et al. 2006)**.

We used partial correlation since partial correlations represent direct associations, while standard correlation analyses do not distinguish between indirect and direct associations, thus partial correlations are more likely to represent primary dependencies and causative links between host and pathogen transcripts. Partial correlations are in general much smaller in value than standard correlations (see for instance **(Altenbuchinger, Zacharias, et al. 2019)**). The association between two nodes i and j was defined as

$$r_{ij} = \begin{cases} r_{ij} & \text{if } P_{adjust} \leq 0.05 \\ 0 & \text{otherwise} \end{cases}, \tag{5.1}$$

where, r_ij is the association between genes i and j and P_adjust is the corrected Pvalue for multiple tests. The connectivity (node degree) of a gene is defined as

$$conn_j = \sum_{i=1}^{p} r_{ij}. \tag{5.2}$$

5.2.9 Functional analysis

TopGO R-package v2.42.0 was used for functional category enrichment analysis **(Alexa et al. 2010)** using the human genes from each bacterial sub-network as the target sets of interesting genes and the list of all genes from human genome build GRCh38.p12 downloaded from Ensembl Biomart **(Kinsella et al. 2011)** as the background set. The Biological Process ontology from Gene Ontology was used **(G. O. Consortium 2019)** and Fisher's exact test was selected to calculate the statistical significance of enrichment for the genes of interest.

5.2.10 Inference of host-pathogen gene association networks at the patient level

We used the Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) method to infer the single sample networks **(Kuijjer, Tung, et al. 2019; Jahagirdar and Saccenti 2020a)** from **Chapter 2** *Personalising Metabolomics? A Closer Look at Single Sample Network Inference*. For sample q (containing gene expression profiles for host and pathogen) out of n samples, the corresponding LIONESS single sample network is obtained as

$$E^q = n\Big(E^\alpha - E^{(\alpha-q)}\Big) + E^{(\alpha-q)}, \tag{5.3}$$

where, E^q is the single sample network of the q^{th} sample, E^α is the aggregate network constructed from all the samples and $E^{(\alpha-q)}$ is the network constructed from all samples excluding the q^{th} sample.

The networks have been estimated using the same approach described in *Gene-gene association network inference* section. Estimation was done separately for streptococcal and polymicrobial NSTI. Note that the aggregate network corresponds to the

host-pathogen gene-gene association networks described in the *Gene-gene association network inference* section.

5.2.11 Clustering of patients based on single-sample networks

Each $m \times m$ single-sample network can be reduced to a $\frac{1}{2}m(m-1)\times 1$ vector containing the perturbations (edges) of the host-pathogen gene correlation. We collapsed these vectors in two matrices of size 2556×42 and 2691×25. For each matrix pairwise distances (Euclidean) among samples were calculated, and hierarchical clustering was applied using the Ward linkage method **(Rokach et al. 2005)**.

5.2.12 Selection of most relevant single sample network edges

We defined the relevance A_{ij} of each single-sample network edge (describing the perturbation of the association between host gene i and pathogen gene j) for each group G of patients defined by applying clustering on the single-sample edges as

$$A_{ij} = \sum_{s=1}^{n} e_{ij}^{s}, \tag{5.4}$$

where e_{ij} is the i,j^{th} edge in the single sample network for the q^{th} subject in group G; the sum runs on the 1,2,...,n(G) subjects in group G. Single-sample network edges were then ranked based on per each group. For each group, we retained the 10 most relevant edges.

5.3 Results

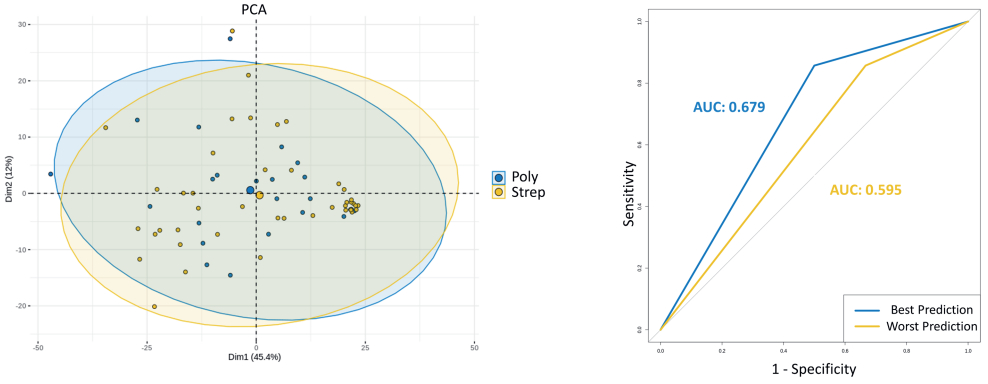


Figure 5.2: **A)** Principal component analysis of the gene expression profiles of streptococcal (n = 42) and poly microbial (n=24) NSTI patients. **B)** Random forest classification of NSTI patient using gene expression profiles. Models are built with cross-validation, and the worst and best model over 100 repetitions are given.

A total of 66 samples were included, comprising 42 monomicrobial *S. pyogenes* NSTIs and 24 polymicrobial cases. The most abundant genera detected across all samples are *Streptococcus*, *Fusobacterium*, *Peptostreptococcus*, *Parvimonas*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Bacteroides* and *Escherichia*. We detected sequences unique to the following species: *S. pyogenes*, *S. dysgalactiae*, *E. coli*, *P. asaccharolytica*, *P. micra* and *Prevotella oris*. The monomicrobial streptococcal NSTI samples were restricted to patient cases with *S. pyogenes* infection. A principal component analysis plot of *S. pyogenes* and polymicrobial NSTI gene expressions and a Random Forest classifier discriminating between them can be found in Figure 5.2.

5.3.1 Host-pathogen gene interaction networks

The interaction networks between human and bacterial genes are shown in figure 5.3. The network specific to monomicrobial *S. pyogenes* NSTI comprises the interaction of 20 human and 24 *S. pyogenes* genes, while the network for polymicrobial NSTI consists of 69 human and 79 bacterial genes.

We observed NSTI type-specific responses in the host, with different sets of human genes highly correlated with bacterial gene expression depending on whether the infection is caused by *S. pyogenes* or by multiple bacteria. We found the polymicrobial correlation network to be divided into subnetworks with genes from three bacterial species (*S. pyogenes*, E.coli and *P. asaccharolytica*) that have a high relative abundance over the samples. While the input gene expression matrix also contained several gene sequences for *P. micra* and *P. oris*, only a single gene interaction with a host gene for each of these species was observed in the resulting networks.

The genes from these association networks were isolated based on species for enrichment analysis. The set of human genes in the *S. pyogenes* monomicrobial network (consisting of *S. pyogenes* genes and associated human genes) were significantly enriched (adjusted P-value <0.05) in GO terms for cytokine production (GO:0080134) and regulation of response to stress (GO:0001816), which include genes coding for the interleukin receptors (IL1R2, IL18R1), CD55 and the heat shock proteins, HSPA5 and HSP90B1 (Fig 5.4). The set of genes in the *S. pyogenes* subnetwork from the polymicrobial samples were significantly enriched in the GO term, cellular response to cytokine (GO:0034097), involving the genes for STAT1, IL18R1, POSTN, CXCL9, CXCL5, demonstrating different responses to mono- versus polymicrobial *S. pyogenes* NSTI.

We observed a strong negative association between the human gene Zinc-Finger Protein 354B (ZFN354B) and three streptococcal genes, sic1 (Q1J9L2), SpyM3_0968 (Q8K763) and MGAS9429_Spy1542 (Q1JK93) in monomicrobial *S. pyogenes*. We also observed strong associations of *S. pyogenes* gene SpyM3_0408 (Q7CFC6) with TATA Box-binding protein- associated factor1D (TAF1D) in polymicrobial infections and with mitochondrially encoded NADH dehydrogenase 4L $MT-ND4L$ in monomicrobial infections. In polymicrobial infections, we found correlations between the human pseudogene Ferritin (FTH1P2) and two *S. pyogenes* genes sagF (Q1JHQ0) and sagG (A2RFD6). *S. pyogenes* genes of known significance are given in table 5.2 and the human genes in 5.3. For all human and bacterial gene interactions and their known functions from Uniprot, we refer you to Supplementary file S2 Chapter 5: Tables S1, S2, S3 and S4.

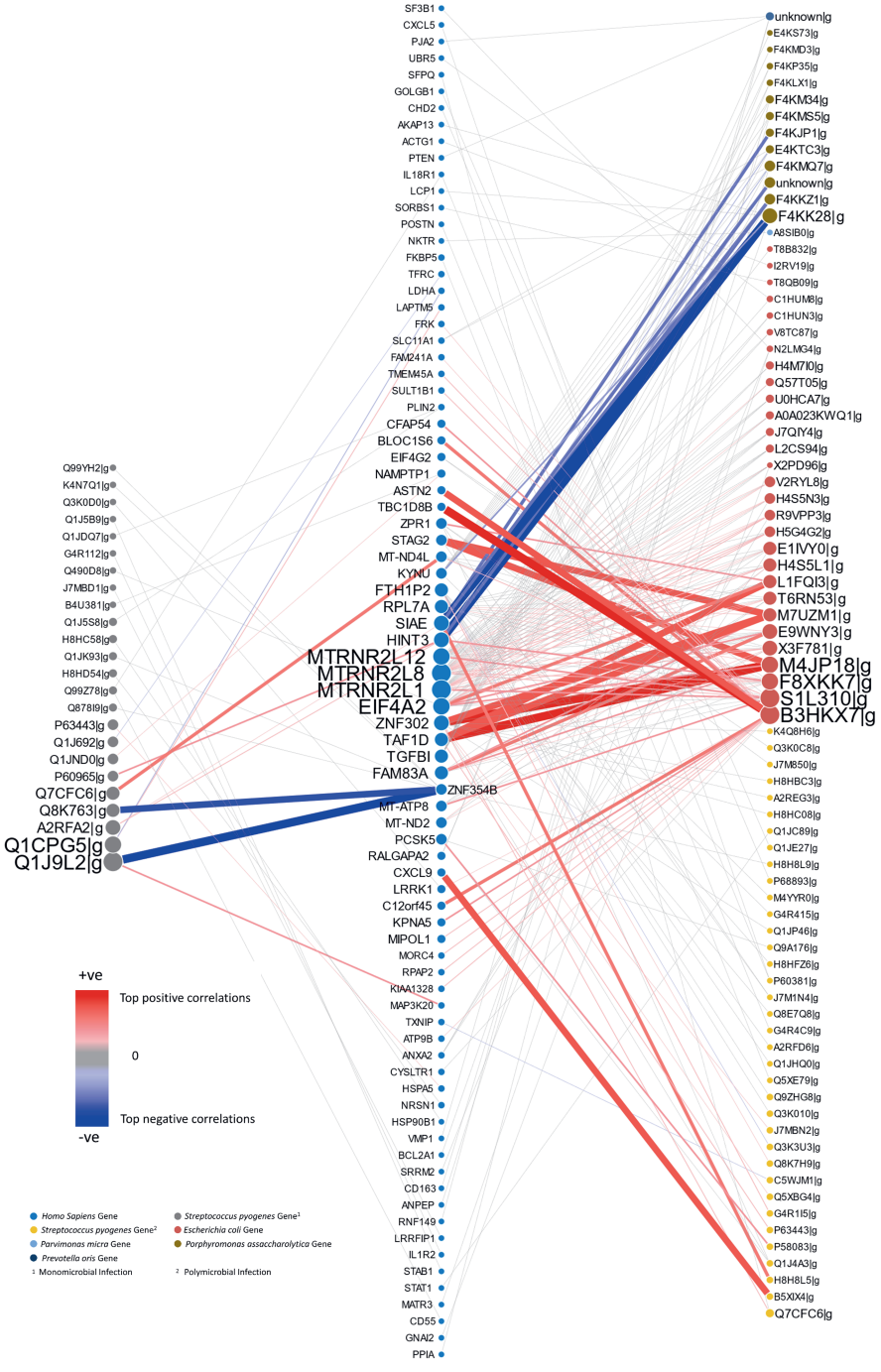

Q99YH2|g
K4N7Q1|g
Q3K0D0|g
Q1J5B9|g
Q1JDQ7|g
G4R112|g
Q490D8|g
J7MBD1|g
B4U381|g
Q1J5S8|g
H8HC58|g
Q1JK93|g
H8HD54|g
Q99Z78|g
Q878I9|g
P63443|g
Q1J692|g
Q1JND0|g
P60965|g
Q7CFC6|g
Q8K763|g
A2RFA2|g
Q1CPG5|g
Q1J9L2|g
SF3B1
CXCL5
PJA2
UBR5
SFPQ
GOLGB1
CHD2
AKAP13
ACTG1
PTEN
IL18R1
LCP1
SORBS1
POSTN
NKTR
FKBP5
TFRC
LDHA
LAPTM5
FRK
SLC11A1
FAM241A
TMEM45A
SULT1B1
PLIN2
CFAP54
BLOC1S6
EIF4G2
NAMPTP1
ASTN2
TBC1D8B
ZPR1
STAG2
MT-ND4L
KYNU
FTH1P2
RPL7A
SIAE
HINT3
MTRNR2L12
MTRNR2L8
MTRNR2L1
EIF4A2
ZNF302
TAF1D
TGFBI
FAM83A
ZNF354B
MT-ATP8
MT-ND2
PCSK5
RALGAPA2
CXCL9
LRRK1
C12orf45
KPNA5
MIPOL1
MORC4
RPAP2
KIAA1328
MAP3K20
TXNIP
ATP9B
ANXA2
CYSLTR1
HSPA5
NRSN1
HSP90B1
VMP1
BCL2A1
SRRM2
CD163
ANPEP
RNF149
LRRFIP1
IL1R2
STAB1
STAT1
MATR3
CD55
GNAI2
PPIA
unknown|g
E4KS73|g
F4KMD3|g
F4KP35|g
F4KLX1|g
F4KM34|g
F4KMS5|g
F4KJP1|g
E4KTC3|g
F4KMQ7|g
unknown|g
F4KKZ1|g
F4KK28|g
A8SIB0|g
T8B832|g
I2RV19|g
T8QB09|g
C1HUM8|g
C1HUN3|g
V8TC87|g
N2LMG4|g
H4M7I0|g
Q57T05|g
U0HCA7|g
A0A023KWQ1|g
J7QIY4|g
L2CS94|g
X2PD96|g
V2RYL8|g
H4S5N3|g
R9VPP3|g
H5G4G2|g
E1IVY0|g
H4S5L1|g
L1FQI3|g
T6RN53|g
M7UZM1|g
E9WNY3|g
X3F781|g
M4JP18|g
F8XKK7|g
S1L310|g
B3HKX7|g
K4Q8H6|g
Q3K0C8|g
J7M850|g
H8HBC3|g
A2REG3|g
H8HC08|g
Q1JC89|g
Q1JE27|g
H8H8L9|g
P68893|g
M4YYR0|g
G4R415|g
Q1JP46|g
Q9A176|g
H8HFZ6|g
P60381|g
J7M1N4|g
Q8E7Q8|g
G4R4C9|g
A2RFD6|g
Q1JHQ0|g
Q5XE79|g
Q9ZHG8|g
Q3K010|g
J7MBN2|g
Q3K3U3|g
Q8K7H9|g
C5WJM1|g
Q5XBG4|g
G4R1I5|g
P63443|g
P58083|g
Q1J4A3|g
H8H8L5|g
B5XIX4|g
Q7CFC6|g
+ve
Top positive correlations
0
Top negative correlations
-ve
Homo Sapiens Gene
Streptococcus pyogenes Gene²
Parvimonas micra Gene
Prevotella oris Gene
Streptococcus pyogenes Gene¹
Escherichia coli Gene
Porphyromonas assaccharolytica Gene
¹ Monomicrobial Infection
² Polymicrobial Infection

Figure 5.3

Figure 5.3: Interactome network of the host-pathogen gene expression profiles derived from Dual RNA-seq of tissue biopsies of NSTI patients. The central column contains human genes, the left column contains *S. pyogenes* genes (i.e. bacterial genes found to be associated with human genes in monomicrobial NSTI), the right column contains genes from several bacterial species (i.e. bacterial genes found to be associated with human genes in polymicrobial NSTI). Nodes are colour coded by bacterial species; the node size is proportional to the node degree (connectivity, i.e. the number of associated genes; see equation 5.2). Red edges indicate a positive partial correlation; blue edges indicate a negative partial correlation; the colour intensity and the edge width are proportional to the magnitude of the partial correlations

5.3.2 Host-pathogen gene association networks at the patient level

We used the patient-specific gene-gene correlations to characterise host-pathogen response at a patient level and to stratify patients based on such responses. Patient clusters based on single-sample network edges are shown in Figure 5.5 for streptococcal (monomicrobial) (A) and polymicrobial (B) NSTI.

In the case of *S. pyogenes* NSTI, we found 6 distinct groups with 4 to 10 patients each, while for the polymicrobial NSTI, we individuated 4 distinct groups with 5 to 7 patients. For each one of these groups, we retrieved the top 10 most relevant host-pathogen gene associations, characterising the particular response to the infection of each patient group. These are shown in table 5.4 and table 5.5 for *S. pyogenes* (monomicrobial) and polymicrobial NSTI, respectively. For the polymicrobial case, the top relevant associations involve genes that were mapped to five bacteria species namely *S. pyogenes*, *E. coli*, *P. asaccharolytica*, *P. micra* and *P. oris*. We were unable to ascertain associations between the groups and clinical outcomes of patients with significant statistical power due to the lack of sufficient samples per group.

5.4 Discussion

We explored the differences in the host-pathogen transcriptional responses at both the population and individual levels. At the global level, we investigated the differences between the interactomes associated with *S. pyogenes* and polymicrobial NSTI, across the entire cohort of NSTI patient tissue samples. The functions associated with the corresponding proteins in the interactome are shown in Figure 5.6. To model the phenotypic heterogeneity observed in the host-pathogen interactions and dynamics at the patient level, we constructed patient-specific interactome networks. Our results provide further insights into the molecular mechanisms underlying the pathophysiology at the tissue site of infection in mono- and polymicrobial NSTIs albeit with some limitations.

5.4.1 Host-pathogen interactome for streptococcal NSTI

In the *S. pyogenes* monomicrobial interaction network, the most enriched GO categories for the human genes in the network were cytokine production (GO:0080134) and regulation of response to stress (GO:0001816) suggesting immune system defensive mechanisms in the host response to changes in expression of specific streptococ-

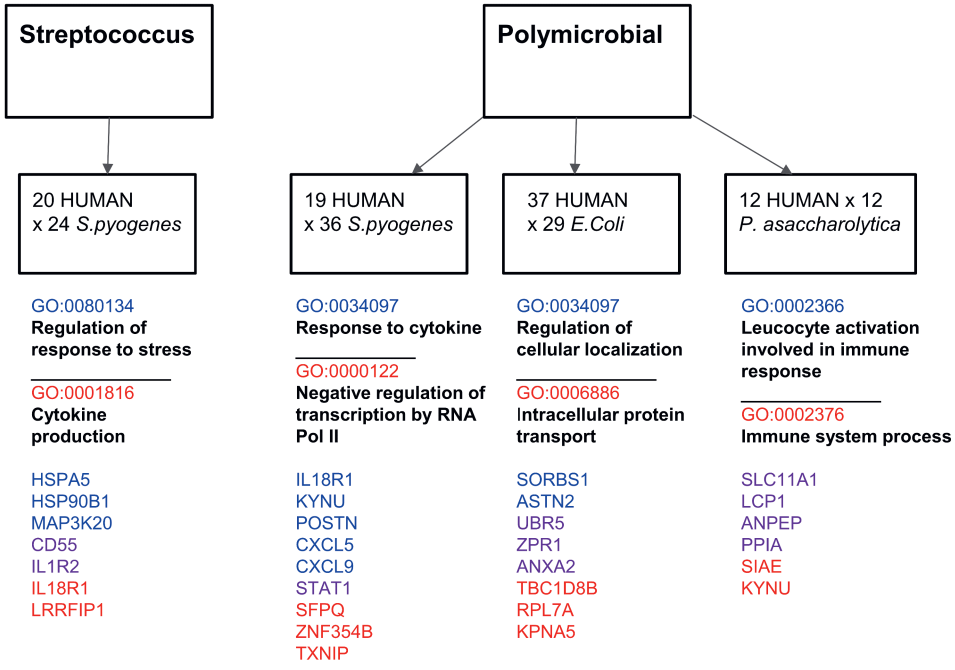


Figure 5.4: GO (Gene Ontology) enrichment analysis of the human and bacterial genes found to be associated and visualised in Fig. 2. Results are given for monomicrobial streptococcal (*S. pyogenes*) and for the polymicrobial NSTI. Gene names listed in blue indicate genes annotated by the GO term also indicated in blue, and gene names in red are annotated by the GO term in red. Genes given in purple colour are annotated by both GO terms. A different type of infection elicits different response patterns in the host

cal genes. The host heat shock protein (HSPA5) is associated with streptococcal acyl carrier protein (P63443, acpP) which is involved in fatty acid biosynthesis in lipid metabolism. Eraso et al. have demonstrated the selection of mutations in the fabT gene, another gene involved in fatty acid biosynthesis during necrotising myositis infections in a non-human primate model **(Eraso et al. 2016)**.

Q1J9L2 is 90% identical to Sic1 **(Merle et al. 2015)**. Sic has several different mechanisms of actions, interference with complement and other host defences, and has been proposed to play a significant role in streptococcal infections **(Westman et al. 2018)**. It was recently shown that the Sic protein from M1 *S. pyogenes*, a type over-represented among severe invasive cases of NSTI **(Bruun, Rath, et al. 2021)**, interacts with TLR2 resulting in release of proinflammatory cytokines **(Neumann et al. 2021)**. A study by Kachroo et al. **(Kachroo et al. 2020)** revealed a positive correlation between sic and genes in table 5.3. Overview of the most relevant human genes obtained from the analysis of host-pathogen gene-gene association networks Different genes were associated to different NSTI types, mono- and polymicrobial involved in the host immune response and inflammation when they examined the dual RNA-seq transcriptomes of *S. pyogenes* and host skeletal muscle from infected

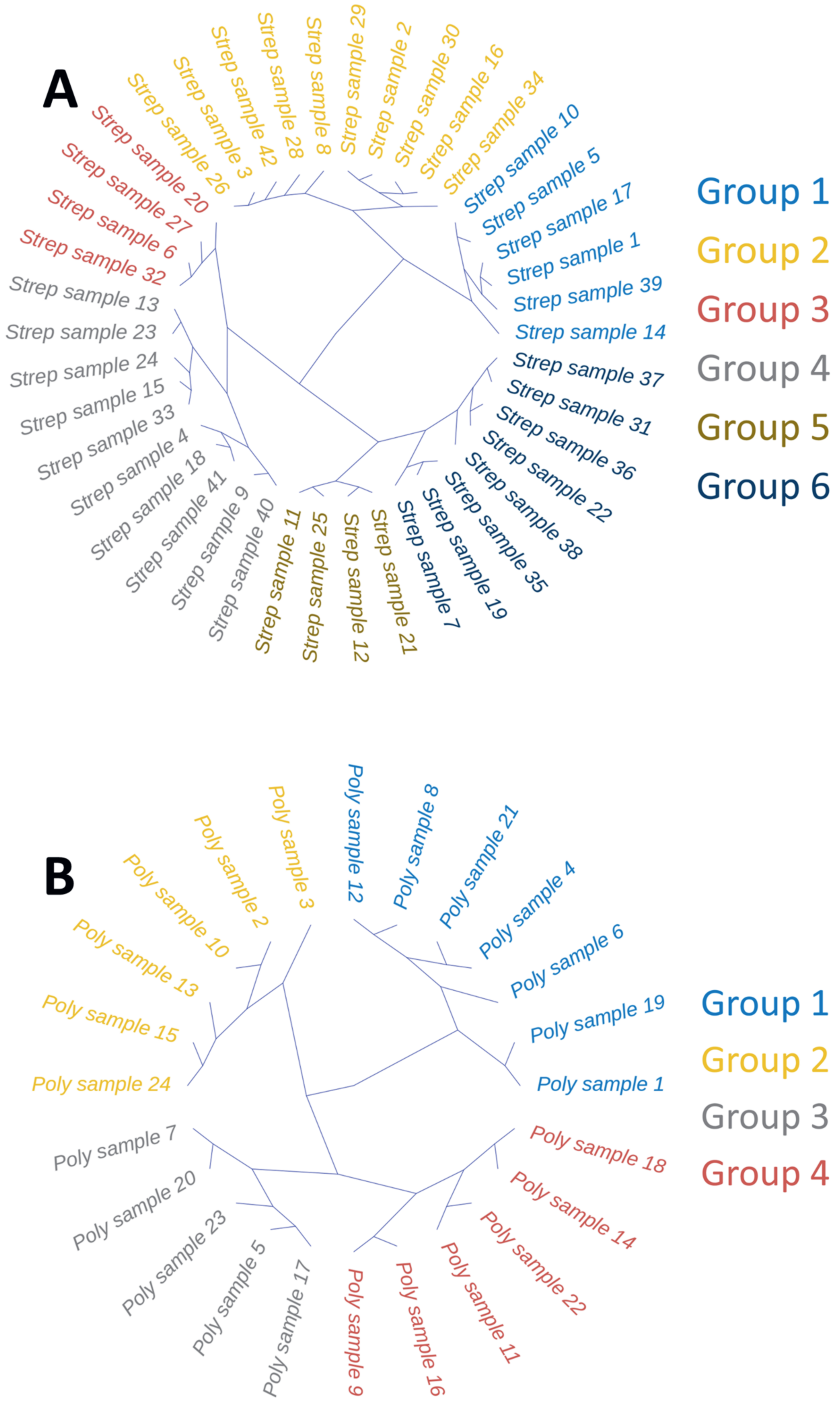

A
Strep sample 29
Strep sample 2
Strep sample 30
Strep sample 16
Strep sample 34
Strep sample 10
Strep sample 5
Strep sample 17
Strep sample 1
Strep sample 39
Strep sample 14
Strep sample 37
Strep sample 31
Strep sample 36
Strep sample 22
Strep sample 38
Strep sample 35
Strep sample 19
Strep sample 7
Strep sample 21
Strep sample 12
Strep sample 25
Strep sample 11
Strep sample 40
Strep sample 9
Strep sample 41
Strep sample 18
Strep sample 4
Strep sample 33
Strep sample 15
Strep sample 24
Strep sample 23
Strep sample 13
Strep sample 32
Strep sample 6
Strep sample 27
Strep sample 20
Strep sample 26
Strep sample 3
Strep sample 42
Strep sample 28
Strep sample 8
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
B
Poly sample 12
Poly sample 8
Poly sample 21
Poly sample 4
Poly sample 6
Poly sample 19
Poly sample 1
Poly sample 18
Poly sample 14
Poly sample 22
Poly sample 11
Poly sample 16
Poly sample 9
Poly sample 17
Poly sample 5
Poly sample 23
Poly sample 20
Poly sample 7
Poly sample 24
Poly sample 15
Poly sample 13
Poly sample 10
Poly sample 2
Poly sample 3
Group 1
Group 2
Group 3
Group 4

Figure 5.5

Table 5.2: Overview of the most relevant *S. pyogenes* genes obtained from the analysis of host-pathogen gene-gene association networks. *S. pyogenes* could be identified in both NSTI types, mono- and poly-microbial.

Gene	Name	NSTI type	Description	Reference
Q1J9L2	Sic1	Mono	Complement inhibitor protein	**(Merle et al. 2015)**
Q1JK93	MGAS9429_Spy1542	Mono	Upstream to gene encoding EndoS (modification of IgG antibodies)	**(Collin et al. 2001)**
H8HD54	MGAS10270_Spy1608	Mono	Downstream to gene encoding EndoS (modification of IgG antibodies)	**(Collin et al. 2001)**
J7MBD1	M1GAS476_1767	Mono	Fibronectin-binding protein (Fba)	**(Terao, Kawabata, Kunitomo, Nakagawa, et al. 2002)**
Q99Z78	MurA2	Mono	Peptidoglycan biosynthesis pathway	**(Engel et al. 2013)**
A2RGM6	STAB902_09315	Mono	Mediator of bacterial signal transduction	**(Bernish et al. 1999)**
F5U6Q2		Mono	Immunogenic secreted protein (Isp)	**(K. S. McIver et al. 1996; U. Consortium 2019)**
P0C0H1	HasA	Mono	Hyaluronic acid capsule (important virulence factor)	**(DeAngelis et al. 1994)**
Q7CFC6	SpyM3_0408	Mono/Poly	Part of ScfAB-operon	**(Le Breton et al. 2017)**
Q9ZHG8	Lbp	Poly	Adhesion to epithelial cells	**(Terao, Kawabata, Kunitomo, Nakagawa, et al. 2002)**
Q1JHQ0	SagF	Poly	Part of the genes that encode for Streptolysin S-operon	**(Shumba et al. 2019)**
A2RFD6	SagG	Poly	Part of the genes that encode for Streptolysin S-operon	**(Shumba et al. 2019)**

Figure 5.5: Hierarchical clustering of single sample networks, i.e. network derived at the patient level as perturbation networks. A) Clustering of single sample networks from Monomicrobial (*S. pyogenes*) NSTI samples B) Clustering of single sample networks from Polymicrobial NSTI samples

nonhuman primates. In addition, vaccination-induced anti-sic antibodies were effective in bacterial clearance in rabbit, mice, and in an ex vivo whole body assay **(Tan et al. 2021)**.

Sic has been shown to bind to extracellular histones, a group of danger signals released during necrotising tissue damage. The aggregates formed from this interaction have been shown both in vitro and in co-localised biopsies from NSTIs resulting in the neutralisation of host antimicrobial activity **(Westman et al. 2018)**. The study by Frick et al. showed that Sic enhances bacterial survival in an animal model of subcutaneous infection **(Frick et al. 2018)**. The increase in the expression of Sic (Q1J9L2) and its role in the inhibition of complement, accompanied by a downregulation of ZFN354B may be a Streptococcal strategy to evade the host innate immune response.

Different bacteria are known to target steps of host gene expression during pathogenic invasion, potentially as a mechanism to modify the expression of inflammatory genes **(Denzer et al. 2020)**. The ZFN354B gene is also negatively correlated with the gene MGAS9429_Spy1542 (Q1JK93). Although, the sequence of this gene

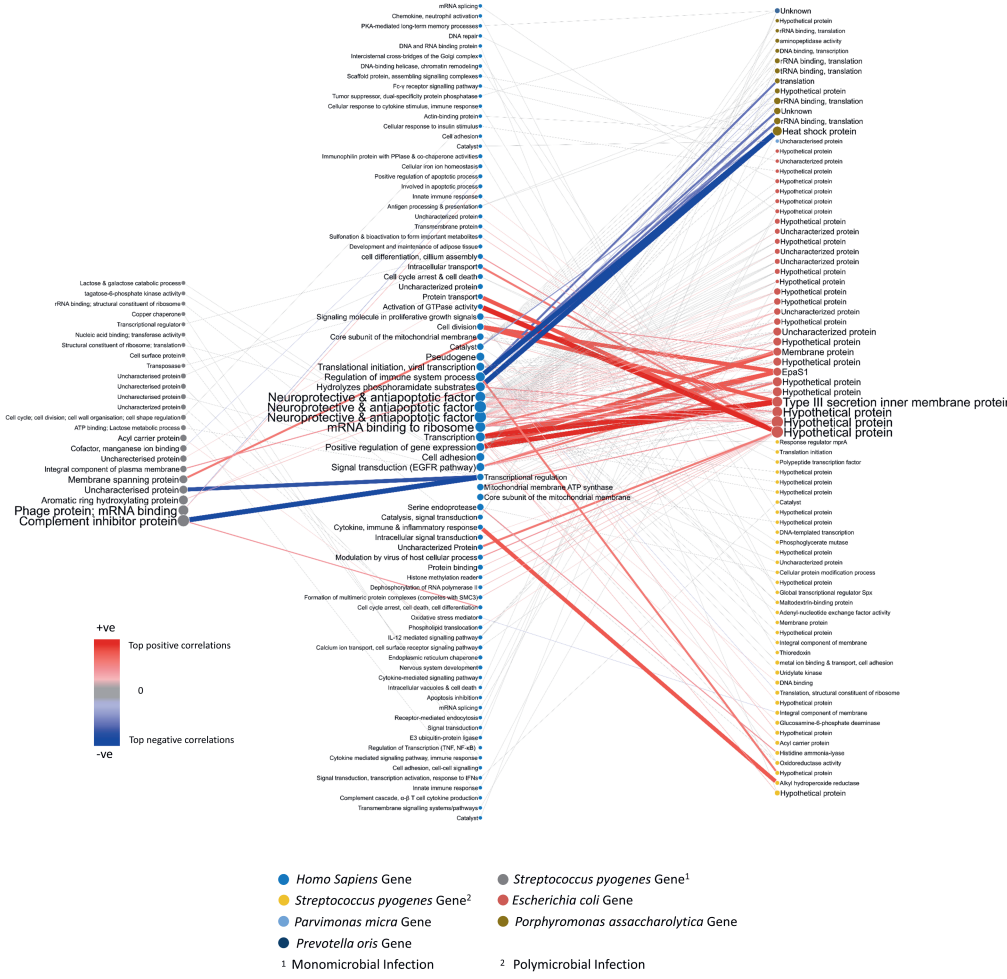


Figure 5.6: Figure of protein functions associated to the host-pathogen gene expression profiles derived from Dual RNA-seq of tissue biopsies of NSTI patients represented by the interactome network given in Fig 5.3. The central column contains human genes, the left column contains *S. pyogenes* genes (i.e. bacterial genes found to be associated with human genes in monomicrobial NSTI), the right column contains genes from several bacterial species (i.e. bacterial genes found to be associated with human genes in polymicrobial NSTI). Nodes are colour coded by bacterial species. The node size is proportional to the node degree (connectivity, i.e. the number of associated genes; see equation 5.2). Red edges indicate positive partial correlation; blue edges indicate negative partial correlation; the colour intensity and the edge width are proportional to the value of the partial correlations

Table 5.3: Overview of the most relevant human genes obtained from the analysis of host-pathogen gene-gene association networks. Different genes were associated to different NSTI types, mono- and polymicrobial.

Gene	Name	NSTI type	Description	Reference
ZFN354B	Zinc-finger protein	Mono	Transcription regulation	**(U. Consortium 2019)**
LRRFIP1	GC-binding factor 2	Mono	Regulation of TNF expression	**(U. Consortium 2019)**
CD55	CD55	Mono	Complement decay-accelerating factor	**(U. Consortium 2019)**
MT-ND4L	Mitochondrially encoded NADH	Mono	Catalyses electron transfer from NADH	**(U. Consortium 2019)**
COL3A1	Collagens	Mono	Structural proteins in the ECM	**(B. Singh et al. 2012; Kuivaniemi et al. 2019)**
COL5A1	Collagens	Mono	Structural proteins in the ECM	**(Greenspan et al. 1992)**
COL6A2	Collagens	Mono	Structural proteins in the ECM	**(Fitzgerald et al. 2008; Glavey et al. 2017)**
FTH1P2	Ferritin	Poly	Intracellular iron storage	**(Di Sanzo et al. 2020)**
KYNU	Kynureninase	Poly	Biosynthesis of NAD co-factors	
CXCL5	C-X-C motif chemokines	Poly	Important role in inflammation	**(M. S. Chang et al. 1994; Erdel et al. 1998; Tokunaga et al. 2018)**
CXCL9	C-X-C motif chemokines	Poly	Important role in inflammation	**(M. S. Chang et al. 1994; Erdel et al. 1998; Tokunaga et al. 2018)**
TGFBI	Kerato-Epithelin	Poly	Cell adhesion & ECM organisation	**(U. Consortium 2019)**
SLC11A1	Solute carrier family 11	Poly	Iron metabolism & host resistance	**(U. Consortium 2019)**
TAF1D	TBP-associated factor 1D	Poly	Component of transcription factor complex	**(U. Consortium 2019)**
TMSB4X	Thymosin Beta - 4	Poly	Organisation of cytoskeleton and actin monomers	**(U. Consortium 2019)**

is unannotated, it is located immediately upstream to the gene encoding EndoS. The gene immediately downstream from the gene encoding EndoS, MGAS10270_Spy1608 (H8HD54), is also found in our analysis correlated with the human gene Neurensin (NRSN1). Even though the exact function of these two gene sequences is unclear, it should be noted that the protein EndoS displays endoglycosidase activity on immunoglobulin G (IgG) by hydrolysing the chitobiose core of the asparagine-linked glycan. EndoS modification of IgG antibodies results in impaired Fcdependent effector function involved in phagocytic killing and elimination of antibody-antigen complexes from circulation **(Collin et al. 2001; Naegeli et al. 2019)**. The strong connectivity of Sic, the EndoS-region, and others in the monomicrobial NSTI networks underlines the importance of the immune evasive strategies in *S. pyogenes* infections. Interestingly, we find that the transcription of both these virulence factors is abated by clindamycin, lending support to contemporary guidelines advocating adjunctive clindamycin treatment in streptococcal NSTIs **(Sartelli et al. 2018)**.

LRRFIP1 gene was found associated with the *S. pyogenes* gene M1GAS476_1767 (J7MBD1) encoding Fba. Fba is a cell-wall-anchoring, surface-located protein that is found in M-type 1, 2, 4, 22, 28 and 49. These M-types constitute 55 out of the 95 sequenced isolates in the INFECT study **(Thänert et al. 2019)**. On studying the effects of Fba in relation to the bacterial invasion of and adhesion to HEp-2 cells, Terao et al. inferred that the presence of both Fba and M-protein are required for the most efficient bacterial adhesion and invasion. In addition, the report showed that a Fba mutant displayed lower mortality in a murine skin infection model **(Terao, Kawabata, Kunitomo, Murakami, et al. 2008)**.

Table 5.4: List of the top associated gene pairs for the four patient groups obtained by hierarchical clustering of single sample networks in the case of monomicrobial (*S. pyogenes*) NSTI. Gene associations highlighted in orange are unique for each group and those highlighted in blue occur in 2 or more groups. Unhighlighted gene associations are present in all groups

Top Associated gene pairs per group			
Group1		**Group2**	
Human	**Strep**	**Human**	**Strep**
C3	J7M6B0|g	C3	J7M6B0|g
CD68	C0M7I0|g	CD68	C0M7I0|g
CTSC	C5WFP7|g	COL3A1	A2RGM6|g
ERBIN	C5WGA7|g	COL6A2	P0C0H1|g
FCGR2A	G4R1C9|g	CYB5R4	P0A4G4|g
FNDC3A	S5EK83|g	ERBIN	C5WGA7|g
IRF1	E8QCI7|g	IRF1	E8QCI7|g
IRF1	P66202|g	IRF1	P66202|g
MYL6	E7PVD6|g	MXD1	J7M930|g
TNNC2	C5WGB5|g	TFRC	E7PWE4|g
Group3		**Group4**	
Human	**Strep**	**Human**	**Strep**
CD68	C0M7I0|g	C3	J7M6B0|g
COL5A1	F5U6Q2|g	CD68	C0M7I0|g
COL6A2	P0C0H1|g	CYB5R4	P0A4G4|g
DUSP1	E7PVR9|g	FGF7	U3TN16|g
IRF1	E8QCI7|g	FNDC3A	S5EK83|g
MXD1	J7M930|g	GMFB	C5WFP7|g
PEAK1	B4U516|g	IRF1	E7PVD6|g
PSAP	Q5X9P6|g	IRF1	E8QCI7|g
RPL30	G4R1C9|g	MXD1	J7M930|g
TNIP1	A2RGM6|g	MYL6	C5WFP7|g
Group5		**Group6**	
Human	**Strep**	**Human**	**Strep**
CD68	C0M7I0|g	ATP2A2	K4Q9Z5|g
COL6A2	P0C0H1|g	CD68	C0M7I0|g
CYB5R4	P0A4G4|g	COL6A2	P0C0H1|g
ERO1A	C5WGA7|g	CYB5R4	P0A4G4|g
GMFB	C5WFP7|g	GMFB	C5WFP7|g
IRF1	E7PVD6|g	MXD1	J7M930|g
MXD1	J7M930|g	PEAK1	B4U516|g
PLCG2	E7PVR9|g	PLCG2	E7PVR9|g
PSAP	Q5X9P6|g	RPL30	G4R1C9|g
RYR1	Q3K1U4|g	TFRC	E7PWE4|g

Table 5.5: List of the top associated gene pairs for the four patient groups obtained by hierarchical clustering of single sample networks in the case of polymicrobial NSTI. Gene associations highlighted in orange are unique for each group and those highlighted in blue occur in 2 or more groups. Unhighlighted gene associations are present in all groups.

Top Associated gene pairs per group			
Group1		**Group2**	
Human	**Bacteria**	**Human**	**Bacteria**
APOL6	E1IVY0|g_E.coli	APOL6	E1IVY0|g_E.coli
CFAP54	X2PD96|g_E.coli	BLOC1S6	A8SND5|g_P.micra
CLU	J7M1A0|g_S.pyogenes	COL15A1	A8SNB9|g_P.micra
COL15A1	A8SNB9|g_P.micra	COL15A1	A8SND5|g_P.micra
COL15A1	A8SND5|g_P.micra	FCGR3A	A8SJ69|g_P.micra
COL1A2	A8SJ69|g_P.micra	FRK	A8SJQ2|g_P.micra
DDX60L	L2CS94|g_E.coli	MGEA5	E1IVY0|g_E.coli
MORC4	A1AII0|g_E.coli	POSTN	F4KLX1|g_P.asaccharolytica
NBPF19	A8SJZ5|g_P.micra	SH3GLB1	V8TC87|g_E.coli
POSTN	F4KLX1|g_P.asaccharolytica	TBL1XR1	J7QIY4|g_E.coli
Group3		**Group4**	
Human	**Bacteria**	**Human**	**Bacteria**
AKAP13	A1AII0|g_E.coli	APOL6	E1IVY0|g_E.coli
APOL6	E1IVY0|g_E.coli	CD177	E1IVY0|g_E.coli
BLOC1S6	A8SND5|g_P.micra	COL15A1	A8SNB9|g_P.micra
CD177	E1IVY0|g_E.coli	COL15A1	A8SND5|g_P.micra
COL15A1	A8SNB9|g_P.micra	COL1A2	A8SJ69|g_P.micra
COL15A1	A8SND5|g_P.micra	MGEA5	E1IVY0|g_E.coli
HSPA5	F5TAH1|g_P.micra	POSTN	F4KLX1|g_P.asaccharolytica
POSTN	F4KLX1|g_P.asaccharolytica	SH3GLB1	V8TC87|g_E.coli
STAT1	S1L310|g_E.coli	STAT1	S1L310|g_E.coli
SYNPO2	Q5X9R3|g_S.pyogenes	TMSB4X	A8SND5|g_P.micra

5.4.2 Host-pathogen interactome for polymicrobial NSTI

In the *S. pyogenes* subnetwork from the polymicrobial interaction network, the most enriched GO category for the human genes was GO:0034097 (Response to Cytokine) with 6 genes (STAT1, IL18R1, KYNU, POSTN, CXCL9, CXCL5) out of 20 annotated with this GO term. IL18R1, the IL-18 receptor complex and the chemokine CXCL5 are associated with bacterial Spx which has an important role in growth, general stress protection and biofilm formation in S. aureus **(Pamp et al. 2006)**. Therefore, the transcriptional response to cytokines in the host is associated with a change in the regulation of transcription in *S. pyogenes* that may impact its ability to form biofilm and modulate its response to stress.

The human gene Ferritin heavy chain 1 Pseudogene 2 (FTH1P2) was found associated with three *S. pyogenes* genes lbp (Q9ZHG8), sagF (Q1JHQ0) and sagG (A2RFD6). Although fth1p2 is a pseudogene, studies have recognised it to compete with fth1, the main intracellular iron-storage protein in the cytoplasm **(Di Sanzo et al. 2020)**. The study by Terao et al. on adhesion of *S. pyogenes* to the HEp-2 cells showed that the absence of Lbp significantly lowered the efficiency of adhesion to epithelial cells **(Terao, Kawabata, Kunitomo, Nakagawa, et al. 2002)**. More studies have indicated that the primary function of Lbp is as a zinc-scavenger and referred to the gene as adcA **(Tedde et al. 2016; Bayle et al. 2011)**. The genes sagF and sagG are two out of the nine genes that form the Streptolysin S-operon. The toxin Streptolysin S has been shown to be responsible for the Beta-haemolysis observed by *S. pyogenes* and has also been implicated in NSTI pathogenesis **(Shumba et al. 2019; Siemens and Norrby-Teglund 2017; Arad et al. 2011; Chatila et al. 1993)**. The association of these *S. pyogenes* genes to fth1p2 is unclear. Enrichment of immune-related host genes was also found in 6 out of the 12 host genes in the *P. asaccharolytica* subnetwork which are annotated with the GO category - Immune system process (GO:0002376). *E. coli* gene interactions with host genes did not reveal any enrichment of genes with immune-related functions. In this subnetwork, enrichment of host genes in cellular localisation and intracellular protein transport was observed and several of the host genes have roles in transcriptional regulation. The majority of the co-expressed bacterial genes in the *E. coli* subnetwork are uncharacterised proteins.

In the *P. asaccharolytica* subnetwork from the polymicrobial interaction network, the most highly connected human gene is KYNU which is correlated with eight *P. asaccharolytica* genes, five of which code for ribosomal proteins (rpmF, rplU, rpsR, rplF, rpsG), one is the small heat shock protein Hsp20 (encoded by Poras_0808) and two are proteins of unknown function **(Santos et al. 2015)**. Heat shock proteins are chaperones that can interfere with the uncontrolled protein unfolding that occurs under stress, such as the immune response of the host **(Colaco et al. 2013)**. An increased kynurenine pathway activity has been linked to inflammation and immune activation and has been implicated in diverse diseases such as depression and cancer **(D. N. Wilson 2014; Dyer et al. 2010; Sforzini et al. 2019)**.

Expression of the host gene, TGFBI, an important cytokine with broad regulatory role in the immune system was also positively correlated with five ribosomal genes (rplU, rpsR, rplF, rpsG, rpsO) in *P. asaccharolytica*. Two other host genes, SIAE and HINT3, are associated with the ribosomal genes (rplU, rplF) and Hsp20 in this species. SIAE has been functionally associated with autoimmune diseases and preeclampsia **(Tsai et al. 2011)**, and although it has a role in the immune system, it has not pre-

viously been studied in the context of an infection. The host gene, SLC11A1, is also associated with the ribosomal gene rplF. SLC11A1 controls natural resistance to infection with intracellular parasites. Pathogen resistance involves sequestration of Fe(2+) and Mn(2+), cofactors of both prokaryotic and eukaryotic catalases and superoxide dismutase, not only to protect the macrophage against its own generation of reactive oxygen species, but also to deny the cations to the pathogen for the synthesis of its protective enzymes.

5.4.3 Differences between host-pathogen interactions in streptococcal infections in monomicrobial and polymicrobial NSTI

The expression of *S. pyogenes* gene SpyM3_0408 (Q7CFC6) was found to be correlated with the expression of human genes in both monomicrobial and polymicrobial infections. In monomicrobial infections, it was positively correlated with the gene mitochondrially encoded NADH dehydrogenase 4L (MT-ND4L). In polymicrobial infections, it was associated with the gene TATA Box-binding protein-associated factor 1D (TAF1D). The gene SpyM3_0408 (Q7CFC6) itself is transcribed as part of the three-gene scfAB-operon. It was hypothesised by Breton et al. that the scfAB-operon plays an integral role in enhancing adaptation and fitness of *S. pyogenes* during localised skin infection and potentially in the propagation to other deeper tissue in a genome-wide Tn-seq analysis to identify *S. pyogenes* genetic determinants necessary for in vivo fitness using a murine model of skin and soft tissue infection. The scfAB-operon is part of *S. pyogenes* core genome, and the gene encodes for a putative transmembrane protein and was found important at the subepithelial site of infection and for the dissemination of into the bloodstream. Homologues of the scfAB-operon are also found in other pathogenic streptococci and closely related gram-positive pathogens **(Le Breton et al. 2017)**.

Although cytokine gene expression was associated with *S. pyogenes* gene expression in both polymicrobial and monomicrobial infections, the only common host gene involved was IL18R1. This occurrence of IL18R1 in both types of infections stands to reason as the Random Forest (RF) models built from cytokine concentrations in the study by Palma Medina et al. in **Chapter 6** found IL-18 to be the least important cytokine to differentiate between monomicrobial and polymicrobial NSTI **(Medina et al. 2021)**.

The genes representing chemokines CXCL5 and CXCL9 were only found in the *S. pyogenes* subnetwork in polymicrobial infections and not in monomicrobial infections. Thänert et al. found CXCL9, CXCL-10/IP-10 and CXCL11 to be overexpressed in streptococcal NSTI and a recent study analysing cytokines and chemokines in plasma samples from NSTI patients by Palma Medina et al. **(Medina et al. 2021)** in **Chapter 6** found CXCL-10/IP-10 to be a robust biomarker for differentiating between monomicrobial and polymicrobial NSTI infections **(Thänert et al. 2019)**. In accordance with these studies, our analysis shows a difference in the involvement of these chemokines between monomicrobial and polymicrobial *S. pyogenes* response. The nature of this network analysis renders any information regarding over-expression or under-expression unascertainable.

5.4.4 Host-pathogen gene associations at the patient level in streptococcal NSTI

Overall, we observed different associations of genes coding for collagen proteins, consistent with the study by Singh et al. which showed that various human pathogens utilise the proteins found in the extracellular matrix (ECM) such as collagen proteins for the invasion of the host. Invasive pathogens infract the basal lamina and degrade the ECM proteins employing various proteases drafted from the host. Pathogens use these abilities to adhere to and invade the host tissue **(B. Singh et al. 2012)**. Group 2 differs from other groups only for perturbation of the association between COL3A1 and A2RGM6. In group 3, we find the association between collagen type V and the gene (F5U6Q2) encoding for a hypothetical protein with high similarity (91.4%) to an immunogenic secreted protein (Isp). This isp gene is located immediately downstream of the ihk-irr TCS and the gene is highly conserved among *S. pyogenes* strains **(K. S. McIver et al. 1996)**. The function of Isp remains elusive but Kachroo et al. showed that the isp gene contributed to virulence in a necrotising myositis model in non-human primates. Kachroo et al. also postulated the potential role of isp in cell-wall metabolism based on the CHAP domain located at the carboxy terminus **(Kachroo et al. 2020)**.

The P0C0H1 (HasA) is required for hyaluronic acid capsule. The capsule represents an important virulence factor of *S. pyogenes* **(DeAngelis et al. 1994)**. We find this gene associated with collagen VI in groups 2, 3, 5 and 6. Despite Collagen VI having antimicrobial properties **(Abdillahi et al. 2018)**, it has been shown to be a target of adherence by *S. pyogenes* for persistent infections and inducing invasions **(Bober et al. 2010)**. Immunodetection, in vitro binding assays and electron microscopy have shown *S. pyogenes* to have a strong affinity to Collagen VI and evolved adhesins with the ability to mediate interactions between *S. pyogenes* and the host **(Bober et al. 2010)**. Studies have reported the possibility of hyaluronic ncapsule to be a major virulence determinant and is also listed in the PATRIC database as a known virulence factor **(Davis et al. 2020)**. Dinkla et al. **(Dinkla, Rohde, et al. 2003)** demonstrated that in rheumatic fever, *S. pyogenes* had the unique capability to bind and aggregate membrane collagen type IV via the M3 protein or hyaluronic acid capsule in M18 serotype. It has also been shown that the upregulation of hyaluronic acid by *S. pyogenes* in blood is used as a mechanism to mask surface immunogenic determinants and evade antigen-specific antibodies to avoid bacterial death in blood. Dinkla et al. **(Dinkla, Sastalla, et al. 2007)** managed to demonstrate that only the *S. pyogenes* abundant in hyaluronic acid capsule were capable of surviving in human blood containing high levels of antibodies directed against highly conserved bacterial surface proteins. This association may point to a mode of entry and immune evasion by *S. pyogenes*.

5.4.5 Host-pathogen gene associations at the patient level in polymicrobial NSTI

Group 4 differs from other groups only for perturbation of the association between the host gene TMSB4X and the *P. micra* gene A8SND5. The TMSB4X gene plays an important role in the organisation of the cytoskeleton and binds to and sequesters actin

monomers (G actin) inhibiting actin polymerisation. A common target of bacterial pathogens is the host cell actin cytoskeleton, a dynamic system of filaments that is central to shape determination, movement, phagocytosis and intracellular trafficking **(Haglund et al. 2011)**. After invasion, some pathogens remain within a membrane-bound compartment and target actin to subvert membrane trafficking **(Haglund et al. 2011)** by polymerising actin on their surface and use filament assembly to power intracellular actin-based motility, generating actin comet tails that trail the moving bacteria **(Haglund et al. 2011; Welch et al. 2013; Truong et al. 2014)**. The associated *P. micra* gene A8SND5 (rplC) highlights the importance of the expression of ribosomal proteins to increase bacterial protein synthesis indicating a possible remodulation of this interplay between host and pathogen. The association of the *E. coli* gene encoding for protein E1IVY0 and several human genes are present in all 4 groups. The E1IVY0 is an uncharacterised protein with high similarity (86%) to a serine acetyl transferase from Pantoea ananatis, a plant pathogenic gram-negative, facultatively anaerobic gamma proteobacterium which has been shown to help the bacterium survive oxidative stress conditions **(Coutinho et al. 2009; Weller-Stuart et al. 2017)**.

5.4.6 Limitations of the study

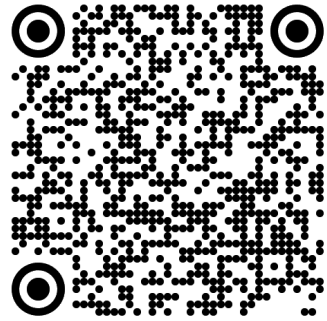
Limitations of this study include the (relatively) small number of samples, heterogeneity among patients with respect to their comorbidities, the time of infection and treatments, along with the differences in biopsy sampling, type of tissue, depth of infection and lack of longitudinal samples. Many of these limitations are unavoidable as the biopsies are obtained at a timing when surgery is clinically indicated and in areas where the tissue pathology warrants surgical removal. However, using important quality aspects such as the collection of biopsies by dedicated teams of clinicians using standardised SOPs, a prospective observational study design similar in all participating clinical centres and the employment of highly stringent statistical approaches strengthen this study.

5.5 Conclusions

Using a systems biology approach to explore host-pathogen interactions in NSTIs, we postulated several data-driven hypotheses which could be further evaluated experimentally. In attempting to elucidate the mechanisms underlying the progression and proliferation of NSTI infections, this study highlights the heterogeneity in the host-pathogen interactomes and strengthens the rationale for a personalised approach in the clinical management of NSTI patients. Furthermore, the identification of pivotal pathogenetic mechanisms is the first step towards identifying novel targets for intervention and expanding our therapeutic armamentarium.

5.6 Data availability

All the data used in this available at
https://doi.org/10.5281/zenodo.5744186.



Publication

5.7 Author contributions

SJ, LM and ES designed the study; NB performed preliminary bioinformatics analysis; SJ and LM performed the analysis; SJ, LM, OO, MS, OH, SS, ANT, VAdP and ES provided advice and interpretation of results; OO, MS, OH, SS, ANT, VAdP and ES provided Material support; SJ and LM visualised the results; SJ, LM and ES wrote the manuscript; OO, MS, SS, VAdP and ES critical revision. All authors reviewed the manuscript and approved the final version.

5.8 Funding

This work was supported by the Center for Innovative Medicine (CIMED) and Region Stockholm (no. 20180058); the Swedish Research Council (2018-02475); the European Union Seventh Framework Programme (FP7/2007-2013) under the grant agreement 305340 (INFECT project); the Swedish Governmental Agency for Innovation Systems (VINNOVA), Innovation Fund Denmark (8114-000005B) and the Research Council of Norway under the frame of NordForsk (project no. 90456, PerAID); the Swedish Research Council, Innovation Fund Denmark (8113-000009B), the Research Council of Norway, the Netherlands Organisation for Health Research and Development (ZonMW) and DLR Federal Ministry of Education and Research, under the frame of ERA PerMed (project 2018-151, PerMIT); and the Swedish Children's Cancer Foundation (TJ2018-0128).

Chapter 6

Chapter

Laura M. Palma Medina[4], Eivind Rath[2*], Sanjeevan Jahagirdar[1*], Trond Bruun[2,3], Martin B. Madsen[5], Kristoffer Stralin[10,11], Christian Unge[10,12], Marco Bo Hansen[5], Per Arnell[13], Michael Nekludov[14], Ole Hyldegaard[5,6], Magda Lourda[4,15], Vitor A. P. Martins dos Santos[7,8], Edoardo Saccenti[1], Steinar Skrede[2,3], Mattias Svensson[4], Anna Norrby-Teglund[4]
***Contributed equally**

Turn to page 377 for author affiliations

This chapter is adapted from:
Palma Medina, L. M., Rath, E*., Jahagirdar, S*., Bruun, T., Madsen, M. B., Strålin, K., Unge, C., Hansen, M. B., Arnell, P., Nekludov, M., Hyldegaard, O., Lourda, M., Santos, V. A. P. M. dos, Saccenti, E., Skrede, S., Svensson, M., & Norrby-Teglund, A. (2021). Discriminatory plasma biomarkers predict specific clinical phenotypes of necrotizing soft-tissue infections. *Journal of Clinical Investigation*, *131*(14). https://doi.org/10.1172/JCI149523
***Contributed equally**

The Race Against Time: Discriminatory Plasma Biomarkers

Abstract

BACKGROUND. Necrotising soft-tissue infections (NSTIs) are rapidly progressing infections frequently complicated by septic shock and associated with high mortality. Early diagnosis is critical for patient outcome, but challenging due to vague initial symptoms. Here, we identified predictive biomarkers for NSTI clinical phenotypes and outcomes using a prospective multicenter NSTI patient cohort.
METHODS. Luminex multiplex assays were used to assess 36 soluble factors in plasma from NSTI patients with positive microbiological cultures (n = 251 and n = 60 in the discovery and validation cohorts, respectively). Control groups for comparative analyses included surgical controls (n = 20), non-NSTI controls (i.e., suspected NSTI with no necrosis detected upon exploratory surgery, *n* = 20), and sepsis patients (*n* = 24).
RESULTS. Thrombomodulin was identified as a unique biomarker for detection of NSTI (AUC, 0.95). A distinct profile discriminating mono- (type II) versus polymicrobial (type I) NSTI types was identified based on differential expression of IL-2, IL-10, IL-22, CXCL-10/IP-10, Fas-Ligand, and MMP-9 (AUC >0.7). While each NSTI type displayed a distinct array of biomarkers predicting septic shock, granulocyte CSF (G-CSF), S100A8, and IL-6 were shared by both types (AUC >0.78). Finally, differential connectivity analysis revealed distinctive networks associated with specific clinical phenotypes.
CONCLUSIONS. This study identifies predictive biomarkers for NSTI clinical phenotypes of potential value for diagnostic, prognostic, and therapeutic approaches in NSTIs.
TRIAL REGISTRATION. ClinicalTrials.gov NCT01790698.

6.1 Introduction

Necrotising soft-tissue infections (NSTIs) are characterised by extensive damage in any layer of the soft-tissue compartment **(Goldstein et al. 2007; Bonne et al. 2017)**. These infections are infrequent but are associated with a significant health burden due to high mortality and risk of severe long-term disability as a consequence of extensive tissue loss or amputations **(A. K. May et al. 2009; Pham et al. 2009; M. B. Madsen, Skrede, et al. 2019)**. The progression of the disease is rapid, and early identification is therefore pivotal for improving the prognosis of affected patients. Currently, the initial diagnosis of NSTI is challenging due to the often vague symptoms during the early stages, a heterogeneous patient group, and lack of specific diagnostic tools **(T. Chan et al. 2008)**, which lead to misdiagnoses of NSTI in many cases **(Goh et al. 2014)**. Still, doctors are advised that in case of NSTI suspicion, patients should be referred for surgical evaluation immediately **(Fernando et al. 2019)**. Previous efforts to improve the diagnosis of NSTIs led to the proposal of the Laboratory Risk Indicator for Narcotising Fasciitis (LRINEC) **(Wong, Khin, et al. 2004)**. However, its utility has been disproven due to low sensitivity **(M. B. Madsen, Skrede, et al. 2019; Fernando et al. 2019; Neeki et al. 2017; Hsiao et al. 2020)**. Therefore, there is still a need for early diagnostic tools facilitating the swift detection of NSTI cases and thereby enabling prompt and adequate treatment **(Peetermans et al. 2020; Saccenti and Svensson 2020)**.

NSTIs are often classified based on aetiology in which 4 types of infections are distinguished; however, the majority of cases consist of types I and II **(D. L. Stevens and Bryant 2017; Giuliano et al. 1977)**. Type I NSTIs are caused by polymicrobial communities working synergistically. This type of infection is the most common type of NSTI, affecting primarily elderly patients and patients with underlying conditions **(Bonne et al. 2017)**. These pathogenic communities include anaerobic and often also aerobic bacteria, including *E. coli* or *Pseudomonas* spp., among others **(Giuliano et al. 1977; Morgan 2010)**. In contrast, type II infections are caused by a single bacterial species, most predominantly by β-hemolytic streptococci, of which *S. pyogenes* (group A Streptococcus (GAS)) is the most common, followed by *S. dysgalactiae* **(Bruun, Rath, et al. 2021)**. This type of NSTI occurs primarily in the extremities of patients that tend to be younger and more often without underlying conditions **(M. B. Madsen, Skrede, et al. 2019; D. L. Stevens and Bryant 2017; Bruun, Rath, et al. 2021)**. Moreover, GAS NSTI cases are often complicated by toxic shock syndrome **(M. B. Madsen, Skrede, et al. 2019; Bruun, Rath, et al. 2021; D. L. Stevens 1995; Study et al. 1997; Darenberg et al. 2007)**. The diversity of microbiological etiologies of these severe infections should translate into different underlying pathogenic mechanisms. In fact, NSTI type–specific host-pathogen interactions were identified by Thänert et al. **(Thänert et al. 2019)** using dual RNA-Seq analyses of tissue biopsies from NSTI patients. This highlighted the possibility of developing diagnostic tools that can contribute to identifying NSTI clinical phenotypes and predicting outcomes, thereby supporting therapeutic strategies that target specific pathogenic mechanisms.

Although these infections are localised in the deep soft tissue, systemic complications are frequently seen and inflammatory mediators have been measured in circulation **(Hansen, Rasmussen, Svensson, et al. 2017)**. This shows the potential of

a diagnostic tool assessing biomarkers in blood, which is advantageous in terms of sampling and options for rapid tests **(T. Chan et al. 2008; Holub et al. 2013)**. In the present study, we explored 36 plasma molecules as potential biomarkers for the detection and characterisation of clinical phenotypes of NSTI using the NSTI patient cohort collected through the prospective multicenter INFECT study **(M. Madsen et al. 2018)**, in which distinct clinical phenotypes involving different comorbidities, localisation, and microbiological aetiology were identified **(M. B. Madsen, Skrede, et al. 2019; Bruun, Rath, et al. 2021)**. We used univariate, multivariate, machine learning, and differential connectivity analyses to identify predictive biomarker sets linked to unique NSTI clinical phenotypes.

6.2 Methods

6.2.1 Patient cohorts

The study is based on clinical data and plasma samples from patients with NSTI (surgically confirmed) enrolled in the multicenter INFECT study. Samples were collected in 5 hospitals in Scandinavia: Blekingesjukhuset (Karlskrona, Sweden), Haukeland University Hospital, Karolinska University Hospital, Righospitalet (Copenhagen, Denmark), and Sahlgrenska University Hospital. Clinical data considered for analyses were recorded at the time of admission to the specialised hospital and were entered into a web-based electronic case report form (eCRF) by trial personnel. Patient characteristics and outcomes for the whole cohort have been reported in Madsen et al. **(M. B. Madsen, Skrede, et al. 2019)**. Of the available plasma samples from the INFECT cohort, 251 samples were considered for the discovery cohort (Figure 6.1). The size of this cohort was limited by technical availability, and samples were selected at random. Due to the lack of other NSTI cohorts with the associated biobank, the validation cohort consisted of a second set of plasma samples from the remaining patient samples from the INFECT study. The size of the validation cohort was determined based on technical availability ($n = 60$). Selection of samples prioritised type II NSTI samples, since only 21 remained available, and then 39 samples from type I NSTI patients were selected at random.

Two additional cohorts of 20 patients each were included as control groups for the discovery cohort. The non-NSTI patient samples were collected during the INFECT study and included patients with suspected NSTI who, after surgical examination, were diagnosed with less severe soft-tissue infections due to lack of necrotic tissue. The surgical controls included patients who had undergone elective surgery at Rigshospitalet for noninfectious conditions and who had no underlying diseases **(Hansen, Rasmussen, Svensson, et al. 2017)**. These 2 control groups were matched in age and sex to the discovery NSTI cohort.

Finally, our study included an additional sepsis cohort of 24 patients (42% septic shock) to determine whether the panel is valid selectively for NSTI cases or would also apply to a broader sepsis patient group. Plasma samples of the sepsis cohort were collected at admission from patients with sepsis at the emergency clinic at the Karolinska University Hospital (Huddinge, Sweden). The size of this cohort was determined by sample availability.

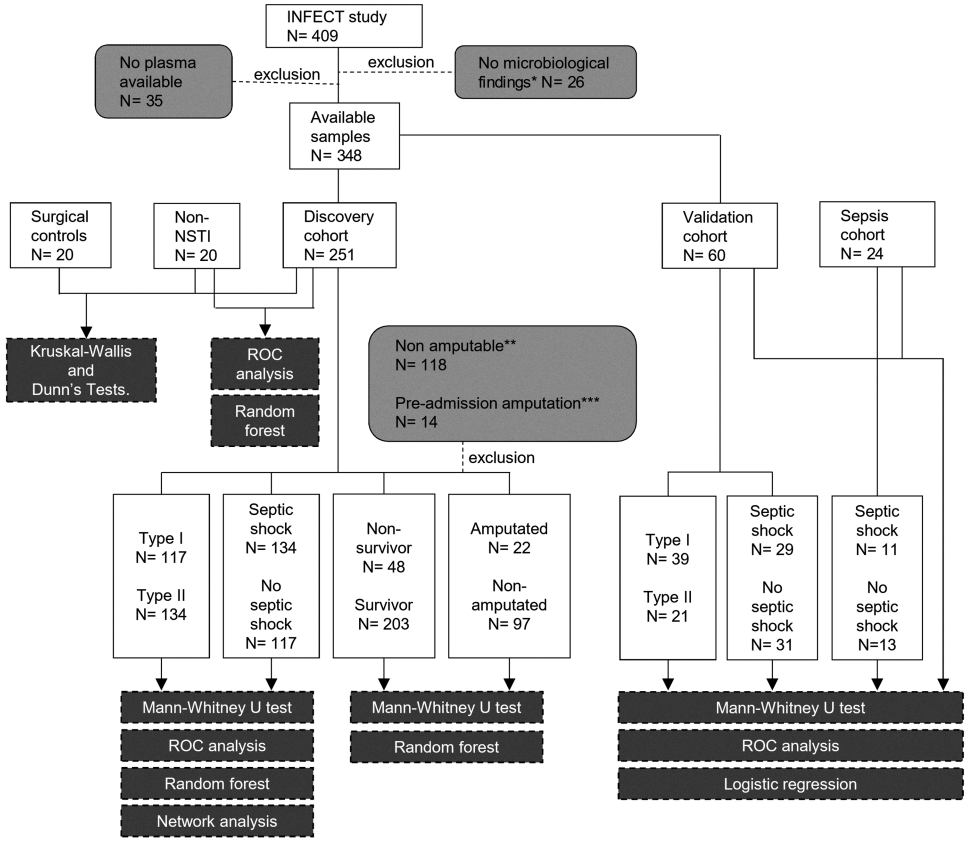


Figure 6.1: Study design. The samples included in each test are displayed inside solid line boxes, light gray boxes show the reasons for exclusion at different stages of the study, and dark gray boxes indicate the specific test applied to the different set of samples. *Plasma samples from the INFECT cohort were excluded from the study if there was no positive microbiological culture in blood or tissue. Samples from patients with NSTI in nonamputable sites (i.e., neck, abdomen, and thorax) (**) or who had undergone amputation before admission (***) were not included for the prediction model for amputation.

6.2.2 Measurement of potential biomarkers in plasma

The plasma samples were prepared from blood collected at admission in EDTA-containing tubes and immediately aliquoted and frozen at –80°C. Concentrations in plasma of the selected list of analytes were determined using the bead-based Luminex multiplex immunoassay. Assays were performed according to the manufacturer's protocol and acquired on a Luminex MAGPIX instrument using xPonent 4.0 software (Luminex). The measurements of the discovery cohort were done in 2 customised multiplex plates of 5 and 32 analytes (R&D Systems). The panel included chemokines (CCL-2/MCP-1, CCL-4/MIP-1β, CCL-5/RANTES, CXCL-8/IL-8, CXCL-10/IP-10), interleukins (IL-1α, IL-1β, IL-2, IL-4, IL-6, IL-10, IL-12p70, IL-13, IL-17A, IL-18, IL-22, IL-36β/IL-1F8), adhesion molecules (E-Selectin, ICAM-1, VCAM-1), matrix metalloproteases (MMP-1, MMP-8, MMP-9), and others (C5/C5a, Collagen-IVα1, Fas-Ligand, Galectin-3, G-CSF, I-α-1/COL1A1, MPO, Pentraxin-3, Resistin, S100A8, S100A9, Thrombomodulin, and TNFα). The initial panel included IL-1RA; however, this analyte was not included in the final analyses due to a high number of out-of-range (OOR) values (>30%). The concentrations of the measured analytes are visualised in Figure 6.2, 6.3, and 6.4.

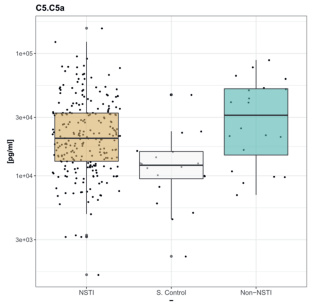
For the validation cohort, only the most robust biomarkers identified in the discovery cohort were assessed. Two panels were measured in customised multiplex plates from R&D Systems (G-CSF, IL-6, S100A8, and Thrombomodulin) and Thermo Fisher (MMP-9, CXCL-10/IP-10, IL-2, IL-10, IL-22, and Fas-Ligand). The results from IL-22 were not included in the final analysis due to a high number of OOR values (>30%).

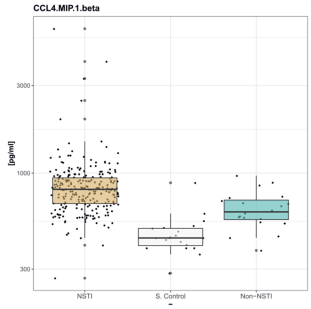
6.2.3 Cell isolation for in vitro validation

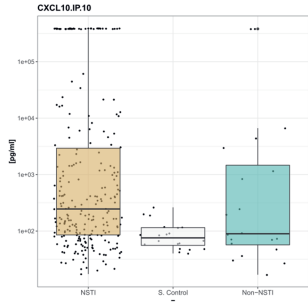
PBMCs were isolated from the peripheral blood of healthy blood donors by Ficoll-Hypaque density gradient centrifugation (Lymphoprep, Axis-Shield) and were resuspended in complete RPMI media (RPMI 1640 [Life Technologies] supplemented with 10% FBS [Sigma-Aldrich], 2 mM l-glutamine [Thermo Fisher Scientific], and 25 mM HEPES [Thermo Fisher Scientific]). The cells were rested overnight at 4°C and were seeded on the day of the experiment at a concentration of 1×10^6 cells/well in a 96-well plate.

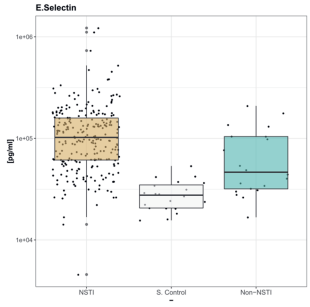
6.2.4 Bacterial strains

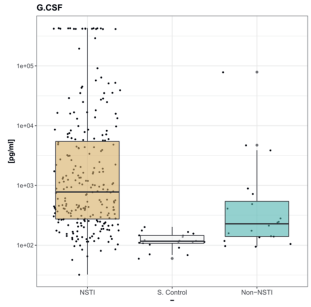
The bacterial strains of GAS, *E. coli*, and *Bacteroides fragilis* are part of the INFECT biobank and were isolated from NSTI patients 2006 (type II NSTI caused by GAS) and 4011 (type I NSTI). *B. fragilis* was cultured inside an Oxoid 2.5 L jar (Thermo Fisher Scientific) with Oxoid AnaeroGen 2.5L sachets (Thermo Fisher Scientific). Bacterial strains were grown from a single colony in brain heart infusion (BHI) broth supplemented with 5% FCS at 37°*C* in an incubator without shaking overnight. After 16 hours, 3 ml of the cultures were collected, and new cultures were inoculated from the ON at an OD600 of 0.05. The bacterial cultures were grown until the late exponential phase (GAS, OD ~ 1; *E. coli*, OD ~ 0.8; and *B. fragilis*, OD ~ 0.6), and 3 ml was collected for further processing.

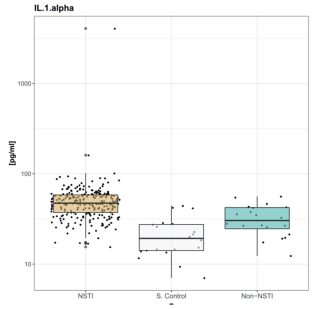

C5.C5a
[pg/ml]
1e+05
3e+04
1e+04
3e+03
NSTI
S. Control
Non−NSTI

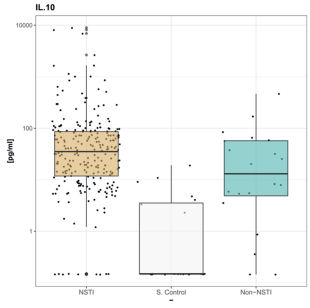

CCL4.MIP.1.beta
[pg/ml]
3000
1000
300
NSTI
S. Control
Non−NSTI

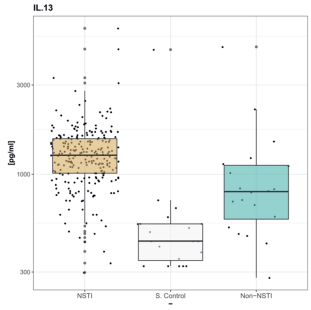

CXCL10.IP.10
[pg/ml]
1e+05
1e+04
1e+03
1e+02
NSTI
S. Control
Non−NSTI

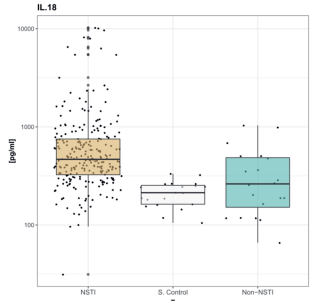

E.Selectin
[pg/ml]
1e+06
1e+05
1e+04
NSTI
S. Control
Non−NSTI

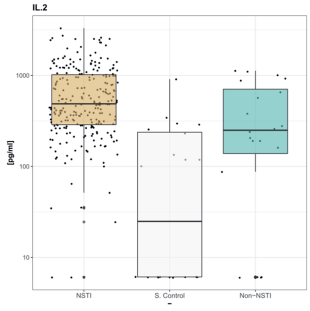

G.CSF
[pg/ml]
1e+05
1e+04
1e+03
1e+02
NSTI
S. Control
Non−NSTI

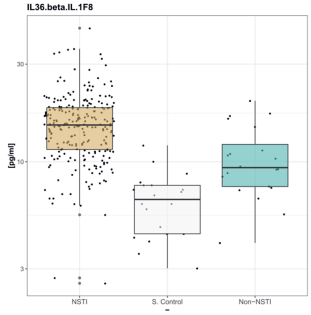

IL.1.alpha
[pg/ml]
1000
100
10
NSTI
S. Control
Non−NSTI

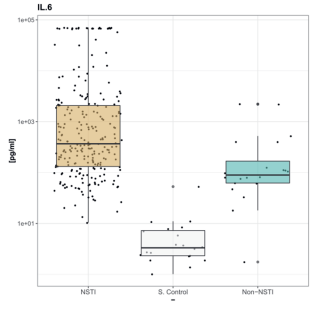

IL.10
[pg/ml]
10000
100
1
NSTI
S. Control
Non−NSTI

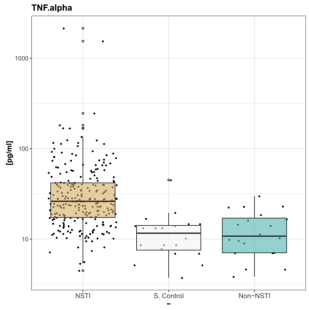

IL.13
[pg/ml]
3000
1000
300
NSTI
S. Control
Non−NSTI

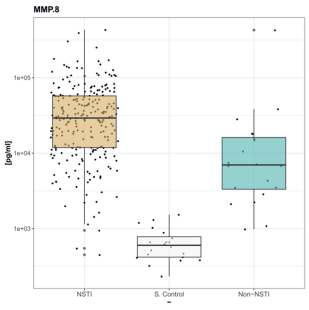

IL.18
[pg/ml]
10000
1000
100
NSTI
S. Control
Non−NSTI

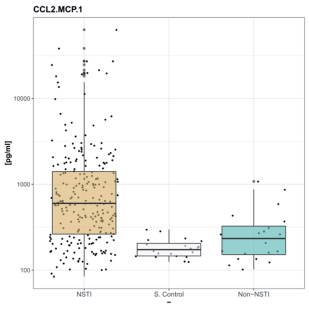

IL.2
[pg/ml]
1000
100
10
NSTI
S. Control
Non−NSTI

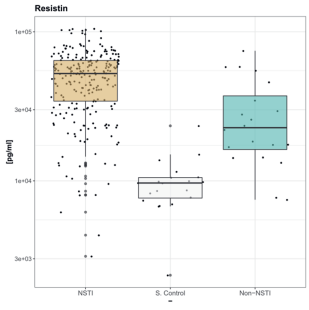

IL36.beta.IL.1F8
[pg/ml]
30
10
3
NSTI
S. Control
Non−NSTI

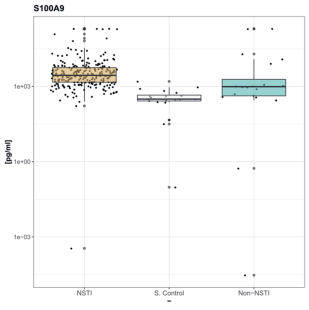

IL.6
[pg/ml]
1e+05
1e+03
1e+01
NSTI
S. Control
Non−NSTI

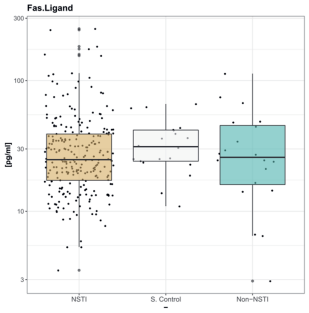

TNF.alpha
[pg/ml]
1000
100
10
NSTI
S. Control
Non−NSTI

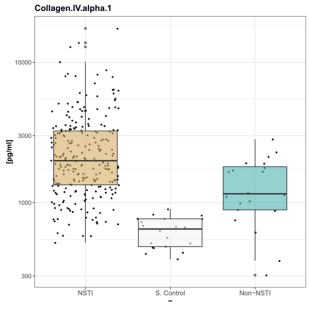

MMP.8
[pg/ml]
1e+05
1e+04
1e+03
NSTI
S. Control
Non−NSTI

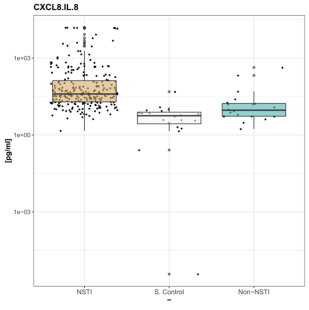

CCL2.MCP.1
[pg/ml]
10000
1000
100
NSTI
S. Control
Non−NSTI

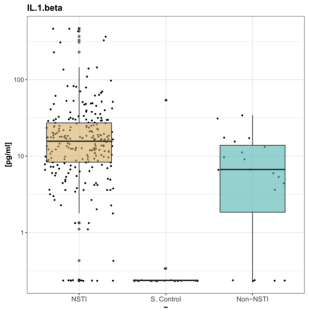

Resistin
[pg/ml]
1e+05
3e+04
1e+04
3e+03
NSTI
S. Control
Non−NSTI

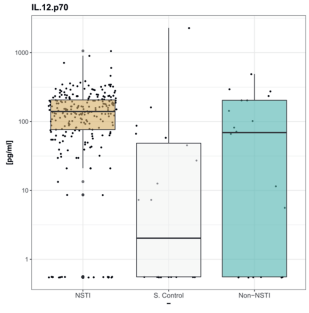

S100A9
[pg/ml]
1e+03
1e+00
1e−03
NSTI
S. Control
Non−NSTI

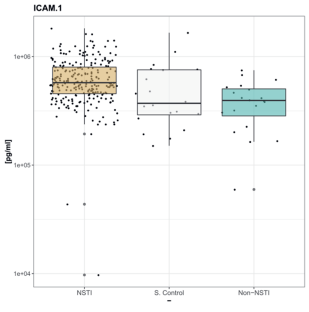

Fas.Ligand
[pg/ml]
300
100
30
10
3
NSTI
S. Control
Non−NSTI

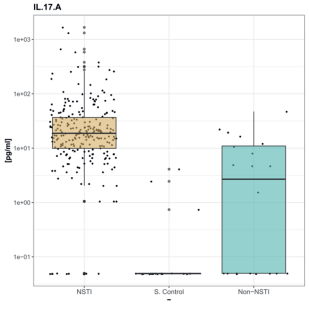

Collagen.IV.alpha.1
[pg/ml]
10000
3000
1000
300
NSTI
S. Control
Non−NSTI


CXCL8.IL.8
[pg/ml]
1e+03
1e+00
1e−03
NSTI
S. Control
Non−NSTI


IL.1.beta
[pg/ml]
100
10
1
NSTI
S. Control
Non−NSTI


IL.12.p70
[pg/ml]
1000
100
10
1
NSTI
S. Control
Non−NSTI


ICAM.1
[pg/ml]
1e+06
1e+05
1e+04
NSTI
S. Control
Non−NSTI


IL.17.A
[pg/ml]
1e+03
1e+02
1e+01
1e+00
1e−01
NSTI
S. Control
Non−NSTI

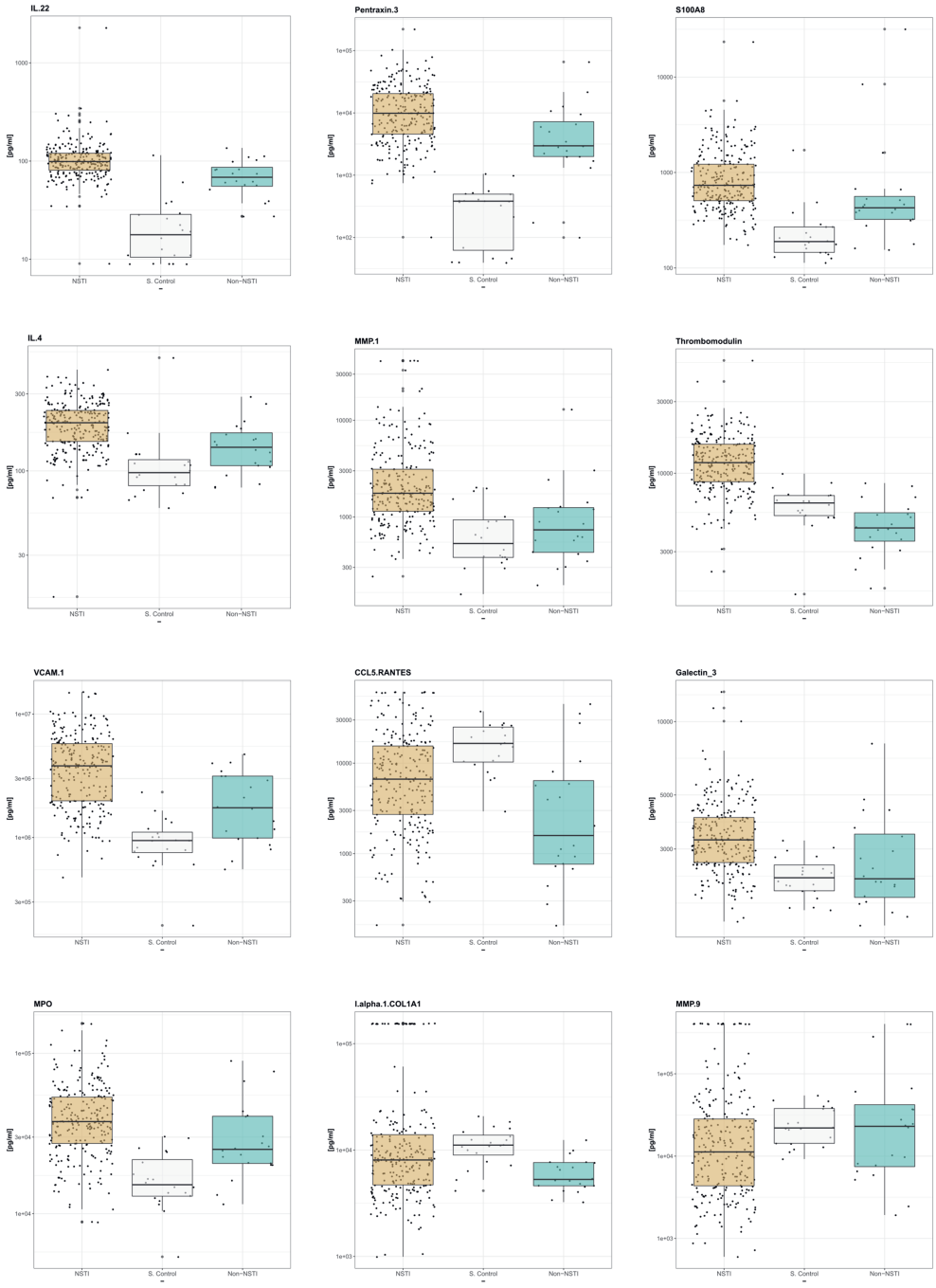

IL.22
Pentraxin.3
S100A8
IL.4
MMP.1
Thrombomodulin
VCAM.1
CCL5.RANTES
Galectin_3
MPO
I.alpha.1.COL1A1
MMP.9
[pg/ml]
NSTI
S. Control
Non−NSTI

Figure 6.4

Figure 6.4: Concentrations of the analytes in the discovery cohort, surgical controls and non-NSTI. The boxplots depict the median and the first and third quartiles, whisker extends to the smallest and largest values or 1.5 times the inter-quartile range at most. NSTI: Necrotising soft tissue infection, S. Control: Surgical controls, Non-NSTI: Suspected NSTI but no necrotic tissue found upon surgical exploration.

6.2.5 PBMC stimulation

Collected bacterial cultures were centrifuged at 1600 g for 10 minutes, and the supernatant was collected and filtered with a 0.2 μm filter. The bacterial pellet was washed and resuspended in 1 ml PBS. The HK sample was prepared by incubation of the resuspended pellet at 75°C for 30 minutes. Stimulation of PBMCs was carried out with a mix of exponential and stationary samples dissolved in RPMI complete. HK samples for stimulation were diluted to an equivalent of a multiplicity of infection of 9, while a 1:200 dilution was used for stimulation with supernatants. The PBMCs were stimulated with samples of GAS, *E. coli*, *B. fragilis*, or a 1:1 mix of the latter two for 24 hours in a 5% CO2 incubator at 37°C. Two control stimulations were included: RPMI media with 10% PBS and RPMI with 10% BHI. Cell culture supernatants were collected by centrifugation at 500 g for 5 minutes. All samples were frozen at -20°C for 16 hours and then transferred to -80°C for long-term storage. In total, 5 independent biological replicates were carried out.

Concentrations of the analytes of interest (CXCL-10/IP-10, Fas-Ligand, IL-2, IL-10, and IL-22) in the cell culture media were determined using the bead-based Luminex multiplex customized plates (R&D Systems) or the IL-22 ELISA Kit (Peprotech) according to the manufacturer's instructions. Prior to measurement, samples were thawed on ice and centrifuged again at 500 g for 5 minutes.

6.2.6 Handling of censored data

Imputation of censored data was only carried out in the discovery cohort. Some of the measured values from the multiplex analyses were found to be below or above the measurable range (OOR; table 6.1), and to generate a complete set of data, these values were imputed. Censored values from cytokines with only one missing value were substituted directly to half the minimal value (left censored) or the maximal plus 20% (right censored). For all other cytokines, the imputation was performed using the method proposed previously **(R. Wei et al. 2018)**. The imputation for cytokines with double-censored data (i.e., both left and right censored) was carried out in 2 sequential imputation steps. First, the data were made left censored by setting the censored values above the range to the maximum observed value and then performing imputation of the left-censored data. Then the right-censored values were set back to be censored and imputed.

6.2.7 Statistics

Significant differences in the clinical data between cohorts or between subsets were tested by Mann-Whitney U test or Fisher's exact test. Wilcoxon's matched-pairs

Table 6.1: Summary of the raw data for the discovery cohort. The numbers of values below (OOR <) or over (OOR >) the range are displayed in the first set of columns. The last two sets of columns display the statistical summary of the data before and after imputation.

	Analyte	OOR data						Raw			Imputed		
		OOR >		OOR <		Total		Median (pg/ml)	Mean (pg/ml)	SD	Median (pg/ml)	Mean (pg/ml)	SD
		#	%	#	%	#	%						
I	CCL-2/MCP-1	0	0%	0	0%	0	0%	438	1840	5570	438	1840	5570
	CCL-4/MIP-1β	0	0%	0	0%	0	0%	768	811	453	768	811	453
	CCL-5/RANTES	10	3%	0	0%	10	3%	6523	10268	10954	6753	11893	13979
	CXCL-8/IL-8	5	2%	1	0%	6	2%	30	321	1408	31	563	2358
	CXCL-10/IP-10	40	13%	0	0%	40	13%	151	1730	5981	197	51320	128964
II	IL-1α	0	0%	0	0%	0	0%	44	58	228	44	58	228
	IL-1β	0	0%	49	16%	49	16%	16	31	62	13	26	58
	IL-2	0	0%	18	6%	18	6%	473	676	577	426	637	581
	IL-4	0	0%	1	0%	1	0%	180	187	71	180	186	72
	IL-6	20	6%	0	0%	20	6%	225	2056	7647	263	6372	18022
	IL-10	1	0%	24	8%	25	8%	31	151	672	25	168	816
	IL-12p70	0	0%	44	14%	44	14%	141	168	179	128	144	176
	IL-13	0	0%	0	0%	0	0%	1159	1220	653	1159	1220	653
	IL-17A	0	0%	36	12%	36	12%	17	48	144	15	42	136
	IL-18	2	1%	0	0%	2	1%	416	699	1092	423	760	1327
	IL-22	0	0%	5	2%	5	2%	94	104	133	93	102	132
	IL-36β/IL-1F8	0	0%	0	0%	0	0%	14	14	6	14	14	6
III	E-Selectin	1	0%	0	0%	1	0%	88779	111397	104435	89170	114977	121780
	ICAM-1	0	0%	0	0%	0	0%	545284	614431	301954	545284	614431	301954
	VCAM-1	0	0%	0	0%	0	0%	3300000	3822748	2774715	3300000	3822748	2774715
IV	MMP-1	3	1%	0	0%	3	1%	1508	2654	4021	1517	3029	5515
	MMP-8	0	0%	0	0%	0	0%	21420	38068	55423	21420	38068	55423
	MMP-9	19	6%	0	0%	19	6%	10661	22758	38480	12209	46289	99356
V	C5/C5a	0	0%	0	0%	0	0%	20311	27081	21269	20311	27081	21269
	Collagen-IVα1	0	0%	0	0%	0	0%	1744	2317	2058	1744	2317	2058
	Fas-Ligand	0	0%	0	0%	0	0%	25	34	31	25	34	31
	Galectin-3	0	0%	0	0%	0	0%	3096	3373	1366	3096	3373	1366
	G-CSF	16	5%	0	0%	16	5%	422	7263	33274	454	28764	97703
	I-α-1/COL1A1	29	9%	0	0%	29	9%	7552	8900	6515	7926	22519	42846
	MPO	1	0%	0	0%	1	0%	34596	39053	22726	34762	39424	23605
	Pentraxin-3	0	0%	2	1%	2	1%	7007	13606	19296	6959	13518	19264
	Resistin	0	0%	0	0%	0	0%	45965	44708	23887	45965	44708	23887
	S100A8	0	0%	0	0%	0	0%	633	1083	2339	633	1083	2339
	S100A9	7	2%	5	2%	12	4%	2197	5455	14986	2213	9707	32129
	Thrombomodulin	0	0%	0	0%	0	0%	10228	11332	5970	10228	11332	5970
	TNFα	0	0%	0	0%	0	0%	23	43	150	23	43	150

signed-rank test was used to test the differences in the in vitro stimulations. Statistical testing of the multiplex results was performed using the Kruskal-Wallis test, followed by Dunn's post hoc test in the case of a 3-group comparison (i.e., NSTI vs. non-NSTI vs. surgical controls) or Mann-Whitney U test for comparison of 2 groups of samples.

To consider unequal group sizes in the discovery cohort, we used a resampling approach to make statistical comparisons. The groups to be compared were made the same size by randomly sampling k samples from each group, where k was chosen to be equal to 90% the size of the smallest group, and then performing statistical testing on these equally sized subgroups. The overall procedure was repeated 10^4 times. We deemed robust and generalisable only those comparisons that were found to be significant in 95% of the runs. The adjustment of P-values was done using the Benjamini-Hochberg adjustment method **(Benjamini et al. 1995)**. Statistical tests for biomarkers linked to the risk of amputation excluded patients with NSTI in non-amputable sites (i.e., neck, abdomen, and thorax) or who had undergone amputation before admission. Statistical comparisons of the results from the validation cohort did not use the resampling method, and the Mann-Whitney U tests were carried out in a standard manner.

6.2.8 ROC analysis

In addition to the statistical comparison of the biomarker levels between different subsets of patients, the diagnostic ability of each marker was tested by ROC analysis. The optimal threshold was selected as the point closest to the top-left part of the plot, which represents perfect sensitivity and specificity. The differences in group sizes in the discovery cohort were also considered for this test, and therefore the resampling methodology explained above also applied to this analysis. The results from the 10^4 curves were assessed by calculating the mean of all outcomes.

6.2.9 RandomForests

RF models **(Breiman 2001)** for the discovery cohort values were built using 10^5 decision trees, and 6 random cytokines were picked at every split selection. To measure the importance of every cytokine in the classification model, mean decrease Gini index was used. Statistical significance was calculated by the means of permutation test using 100 permutations of the original data sets as implemented in the rfPermute package **(Archer et al. 2016)**.

6.2.10 Logistic regression

For the validation cohort, the association between the selected analytes and different outcomes was assessed by calculating odds ratios based on logistic regression analysis. All analytes' concentrations were transformed with log_2 before the generation of the model. The odds ratios were obtained by exponentiation of the model coefficients. Multivariate logistic regressions were performed to correct for sex and age.

6.2.11 Network analyses

Protein association networks were built using the context likelihood of relatedness based on correlation algorithm (PCLRC), which was first introduced to reconstruct metabolite correlation networks and shown to be robust against variation in sample size and noise **(Jahagirdar, Suarez-Diez, et al. 2019; Suarez-Diez et al. 2015)**. In the present study, we used pairwise partial correlation among proteins measured on the different patient groups as a weighted measure of analyte association to reduce the chances of false indirect associations. PCLRC gives the probability of likelihood of occurrence of a relationship between the cytokines. Associations with probability weights of more than 0.95 were retained in the analysis. Cytokine-association networks were built for different patient groups and compared as detailed below.

6.2.12 Differential connectivity analysis

Differential connectivity was used to compare the cytokine association networks of different patient groups and to highlight cytokines whose patterns of association vary. Differential connectivity analysis has been successful in investigating potential molecular mechanisms underlying different conditions in biological systems **(Afzal et al. 2019)**. The connectivity for each node (i.e., protein) in the network is defined as the summation of the absolute values of the weights of all the edges associated

with the given node, thereby accounting for both the number of connections and the weight of those connections. Thus, for i^{th} cytokine, cytokine connectivity X_i is given by the following:

$$X_i = \sum_{j>i} |r_{ij}|, \tag{6.1}$$

where r is the correlation, as defined by the PCLRC algorithm, between cytokines i and j. The differential connectivity (ΔX_i) of the ith cytokine in networks from group 1 (G1) and group 2 (G2) can be given by the following:

$$\Delta X_i = |X_i^{G1} - X_i^{G2}|. \tag{6.2}$$

The statistical significance of the observed differential connectivity for each cytokine was established by using a permutation test **(Afzal et al. 2019; Jahagirdar and Saccenti 2020b)**. The procedure included the independent permutation of the values of every protein repeated 10^3 times with the intention of deriving a probability of the observation in the form of a P-value.

6.2.13 Software

All statistical tests included in this paper were performed in R, version 3.6 **(R Core Team 2013)**, or GraphPad Prism, version 8.2.0, for Windows. Kruskal-Wallis and Mann-Whitney U tests were performed using the R stats package, and the post hoc Dunn's test was performed using the FSA package **(Ogle et al. 2020)**. The ROC tests were performed with the R package pROC **(Robin et al. 2011)**. The logistic regression was perform using glm, and the confidence intervals were obtained with confint, both functions from the package stats. The RF models were built using the R package rfPermute **(Archer et al. 2016)**. The R code for PCLRC is available at the Waganingen University Laboratory of Systems and Synthetic Biology website (www.systemsbiology.nl) under the software tab. Finally, Fisher's exact test for comparison of clinical data and Wilcoxon's matched-pairs signed-rank test for the statistical comparison of in vitro results were performed using GraphPad Prism, version 8.2.0.

6.2.14 Study approval

The multicentre INFECT study is registered at ClinicalTrials.gov (NCT01790698). The study was approved by national ethics committees, including the Regional Ethical Review Board at the Karolinska Institute in Stockholm, Sweden (ethics permits 2012/2110-31/2), the regional Ethical Review Board at the National Committee on Health Research Ethics in Copenhagen (ref. no. 1211709, including amendment 4:61050; regional ethics committee H-2-2014-071), the Regional Ethics Committee in Gothenburg (ref. no. 930-12), and the Regional Ethics Committee in Vest, Norway (ref. no. 2012/2227/REK VEST). Use of sepsis samples included in the study was approved by the Regional Ethical Board in Stockholm (2017/1358-31) All studies were conducted in accordance with the Declaration of Helsinki. All samples were pseudonymised. All patients or their legal guardians provided informed consent prior to enrolment and sample collection.

6.2.15 Data and materials availability

All data associated with this study are available in the main text. The raw data with and without imputed values are available at Dryad (DOI: 10.5061/dryad.f1vhhmgw4; https://datadryad.org/stash/share/zclF2y-NfdaaSAKY-6N02TmWJhd9oONAayDySmCTzy8).

6.3 Results

6.3.1 Study subjects for the discovery cohort

Study subjects were selected from the INFECT patient cohort **(M. B. Madsen, Skrede, et al. 2019; Bruun, Rath, et al. 2021)**. A key aspect of this study was to ensure that the microbiological aetiology was considered, as this influences the clinical phenotypes and the pathogenic mechanisms. For this purpose, only patients with positive microbiological culture in blood or tissue and with plasma collected at the time of enrolment were included in the analysis. Out of the 348 patients in the INFECT cohort with microbiological results and available plasma samples, 251 patients were selected for the discovery cohort (Figure 6.1). These included 117 type I (47%) and 134 type II (53%) NSTI cases, thus obtaining an aetiology distribution that was representative of the original INFECT cohort **(M. B. Madsen, Skrede, et al. 2019)**. Plasma samples from 2 control groups were also included in the analyses: 20 patients with suspected NSTI in whom no necrotic tissue was found upon explorative surgery (non-NSTI controls) and an additional control group of 20 patients who had surgical procedures not related to infection (surgical controls). The latter cohort was matched with the NSTI patients for age and sex. The associated clinical and microbiological data of the patients and controls are shown in table 6.2. The distribution of age, sex, and simplified acute physiology score II (SAPS II) **(Le Gall et al. 1993)** was similar among all NSTI patients regardless of the type of infection.

6.3.2 Biomarkers discriminating NSTI from non-NSTI controls

To pinpoint relevant markers for early detection of NSTI, a customised Luminex multiplex assay including 36 soluble factors involved in inflammatory responses and tissue remodeling was designed. The biomarker profiles in plasma samples from patients were compared with those of controls, and as expected, the highest concentrations of the markers were typically measured in plasma from NSTI patients, followed by the non-NSTI controls, and finally, the non infected surgical controls (Figure 6.52A and Figure 6.4. The use of a stringent statistical analysis allowed the identification of the most robust biomarkers discriminating between the groups. The results revealed that most analytes' concentrations were significantly higher in NSTI samples than in the non-infected surgical controls ($q < 0.05$), whereas only 4 markers, i.e., IL-6, IL-22, MMP-8, and Pentraxin-3, were significantly higher in non-NSTI samples compared with those from the surgical controls ($q < 0.05$). Most relevant from a clinical perspective, comparison between NSTI and non-NSTI cases revealed that only Thrombomodulin differed significantly between these groups ($q < 0.0005$) (Figure 6.5A). The robustness of this protein as a potential biomarker was corroborated

Table 6.2: Clinical characteristics of the discovery patient cohort and controls. Data are shown as mean values and SD or percentages. NA, not applicable; Strep: Streptococcus sp. Significant differences between cohorts were determined by Mann-Whitney U test or Fisher's exact test. AIncludes only infections in extremities (n = 119; type I = 19). B Within 4 weeks before admission for NSTI.

	NSTI			Non-NSTI (n = 20)	Surgical controls (n=20)	P-values		
	All (n = 251)	Type I (n = 117)	Type II (n = 134)			NSTI vs Non-NSTI	NSTI vs S. control	Type I vs Type II
Age (yr)	59 ± 15	59 ± 14	59 ± 15	46 ± 13	59 ± 19	0.0002	0.794	0.999
Male sex	138 (55%)	69 (59%)	69 (51%)	13 (65%)	11 (55%)	0.486	>0.999	0.254
Female sex	113 (45%	48 (41%)	65 (49%)	7 (35%)	9 (45%)	0.486	>0.999	0.254
Septic shock at baseline	134 (53%)	55 (47%)	79 (59%)	6 (30%)	NA	0.061	NA	0.076
Amputation[A]	22 (18%)	5 (26%)	17 (17%)	0	NA	NA	NA	0.343
90-Day mortality	48 (19%)	27 (23%)	21 (16%)	1 (5%)	NA	0.14	NA	0.15
Comorbidities	180 (72%)	90 (77%)	90 (67%)	NA	NA	NA	NA	0.094
Diabetes (type I or II)	60 (24%)	43 (37%)	17 (13%)	2 (10%)	NA	0.266	NA	<0.0001
Cardiovascular disease	101 (40%)	48 (41%)	53 (40%)	4 (20%)	NA	0.095	NA	0.897
Surgery Before NSTI[B]	37 (15 %)	27 (23%)	10 (7%)	4 (20%)	20 (100%)	0.518	<0.0001	0.001
SAPS II	45 ± 16 (10% NA)	44 ± 15 (8% NA)	46 ± 17 (12% NA)	29 ± 15 (30% NA)	NA	<0.0001	NA	0.418
SOFA score at admission	8 ± 4 (4% NA)	8 ± 3 (3% NA)	9 ± 4 (4% NA)	4 ± 3 (20% NA)	NA	<0.0001	NA	0.017
Type I	117(47%)	NA	NA	11 (55%)	NA	NA	NA	NA
Microbiological findings								
GAS	98 (39%)	10 (9%)	88 (66%)	2 (10%)	NA	0.008	NA	<0.0001
Other strep	40 (16%)	21 (18%)	19 (14%)	4 (20%)	NA	0.544	NA	0.49
S. aureus	17 (7%)	10 (9%)	7 (5%)	2 (10%)	NA	0.64	NA	0.324
Clostridium spp.	12 (5%)	6 (5%)	6 (4%)	0 (0%)	NA	>0.999	NA	>0.999
Others	84 (33%)	70 (60%)	14 (10%)	12 (60%)	NA	0.027	NA	<0.0001

by receiver operating characteristic (ROC) analysis with an AUC of 0.95 (specificity, 0.89; sensitivity, 0.92, at a concentration threshold of 7567 pg/ml), outperforming selected clinical markers (Figure 6.6, A and B, and Table 16.3). Next, multivariate analysis using random forest (RF) modeling, including the whole biomarker set and key clinical parameters (i.e., age, sex, sequential organ failure assessment [SOFA] score **(J. .-. Vincent et al. 1996)**, septic shock, NSTI type, WBC, C-reactive protein [CRP] and creatinine), was used to identify biomarkers predictive of NSTI. The analyses identified a set of 5 biomarkers, i.e., IL-17A, Galectin-3, S100A8, S100A9, and Thrombomodulin, that differentiated between NSTI and non-NSTI cases (Figure 6.5B and Table 6.3). Notably, Thrombomodulin was the most robust predictive marker even in the multivariate model. Furthermore, comparison of NSTI patients divided based on early (severe pain, in need of opioids), intermediate (skin bullae or skin bruising), and late (skin purple/black discoloration, skin anesthesia, palpable gas [crepitus] or gas visualised on radiology) signs of NSTI revealed that even patients with only early signs had significantly higher levels of Thrombomodulin than non-NSTI controls (Figure 6.6). Moreover, these levels further increased in patients with intermediate or late signs of NSTI.

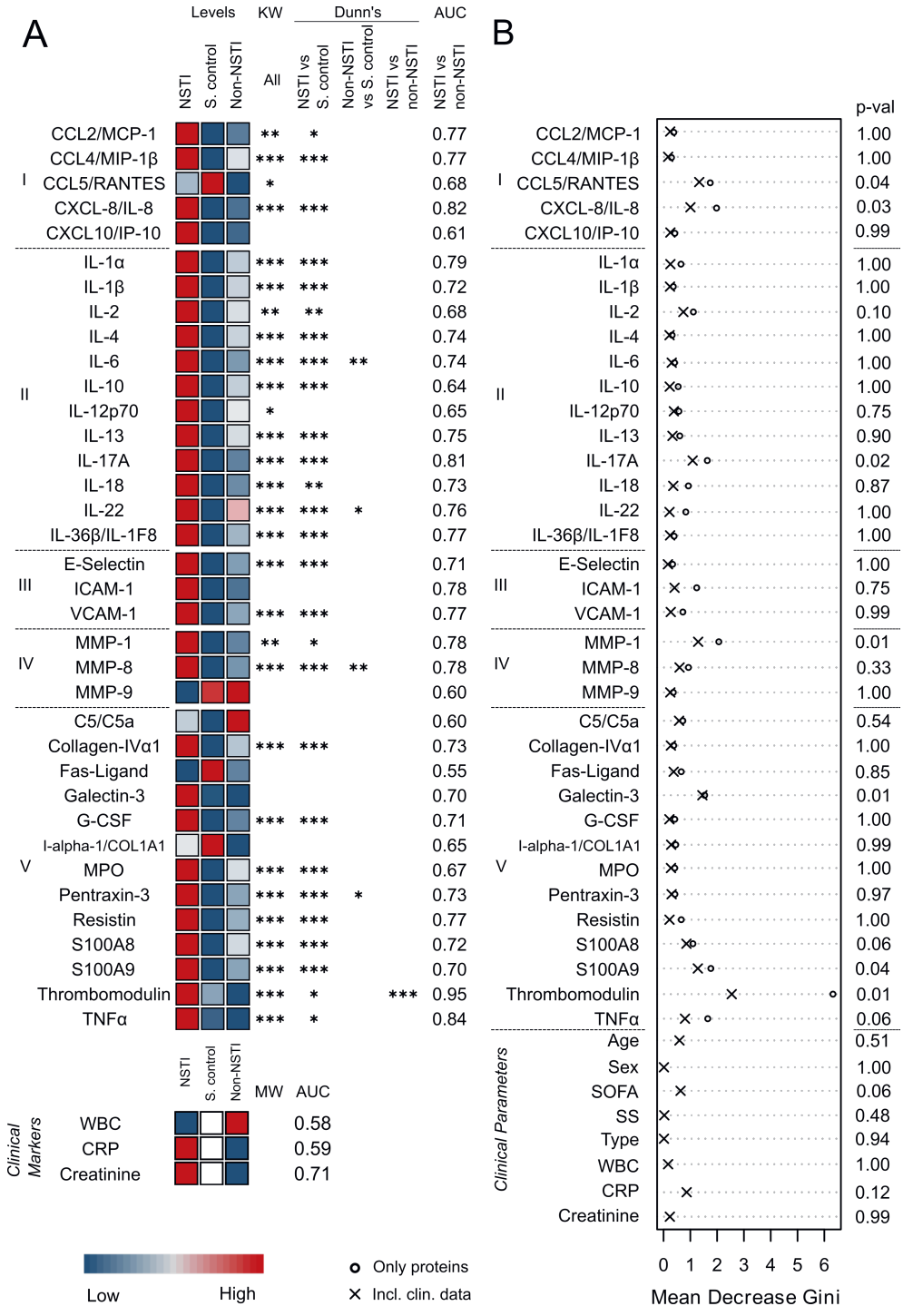

A
Levels
KW
Dunn's
AUC
NSTI
S. control
Non-NSTI
All
NSTI vs S. control
Non-NSTI vs S. control
NSTI vs non-NSTI
NSTI vs non-NSTI
I
II
III
IV
V
CCL2/MCP-1
CCL4/MIP-1β
CCL5/RANTES
CXCL-8/IL-8
CXCL10/IP-10
IL-1α
IL-1β
IL-2
IL-4
IL-6
IL-10
IL-12p70
IL-13
IL-17A
IL-18
IL-22
IL-36β/IL-1F8
E-Selectin
ICAM-1
VCAM-1
MMP-1
MMP-8
MMP-9
C5/C5a
Collagen-IVα1
Fas-Ligand
Galectin-3
G-CSF
I-alpha-1/COL1A1
MPO
Pentraxin-3
Resistin
S100A8
S100A9
Thrombomodulin
TNFα
Clinical Markers
MW
WBC
CRP
Creatinine
Low
High
B
p-val
Clinical Parameters
Age
Sex
SOFA
SS
Type
WBC
CRP
Creatinine
0 1 2 3 4 5 6
Mean Decrease Gini
Only proteins
Incl. clin. data

Figure 6.5

Table 6.3: Results of the ROC analyses and the random forest model discriminating NSTI (discovery cohort) and non-NSTI. The ROC results are shown as the mean values of the iterations. Thresholds are given as pg/ml. The p-values of the random forest results were calculated using 100 permutations of the original dataset. AUC: Area under the curve, CI: Confidence interval, Sn: Sensitivity, Sp: Specificity, MDA: Mean decrease accuracy, MDG: Mean decrease gini; SOFA: Sequential organ failure assessment score, SS: Septic shock, WBC: white blood cells, CRP: C -reactive protein, Type: microbiological classification of NSTI.

Analyte	Receiver operating characteristic					Random Forest							
						Only proteins				with Clinical variables			
	AUC	[95% CI]	Threshold	Sn	Sp	MDA	p-val	MDG	p-val	MDA	p-val	MDG	p-val
CCL-2/MCP-1	0.77	[0.61-0.93]	376.39	0.79	0.71	18.87	1	0.43	1	14.6	1	0.25	1
CCL-4/MIP-1β	0.77	[0.62-0.93]	727.39	0.78	0.72	26.45	1	0.34	1	23.16	1	0.17	1
CCL-5/RANTES	0.68	[0.5-0.86]	3944.64	0.63	0.75	63.37	0.03	1.79	0.13	73.29	0.01	1.33	0.04
CXCL-8/IL-8	0.82	[0.67-0.96]	20.1	0.8	0.8	82.91	0.32	2.02	0.01	56.38	0.44	1	0.03
CXCL-10/IP-10	0.61	[0.42-0.8]	152.17	0.56	0.72	39.35	0.86	0.47	1	29.2	0.89	0.26	0.99
IL-1α	0.79	[0.64-0.94]	40.81	0.76	0.73	44.35	0.99	0.69	0.81	20.37	1	0.26	1
IL-1β	0.72	[0.55-0.89]	11.32	0.69	0.71	32.58	0.99	0.41	1	33.93	0.9	0.24	1
IL-2	0.68	[0.5-0.86]	311.64	0.61	0.77	59.79	0.8	1.15	0.21	31.93	0.95	0.74	0.1
IL-4	0.74	[0.57-0.91]	168.42	0.73	0.71	25.96	1	0.36	1	29.11	0.99	0.22	1
IL-6	0.74	[0.58-0.91]	146.03	0.74	0.77	30.89	1	0.45	1	25.81	0.99	0.31	1
IL-10	0.64	[0.45-0.82]	22.93	0.59	0.72	36.92	0.98	0.59	0.98	20.72	0.95	0.23	1
IL-12p70	0.65	[0.46-0.84]	92.97	0.62	0.74	50.95	0.9	0.61	0.92	52.84	0.66	0.38	0.75
IL-13	0.75	[0.58-0.92]	1012.96	0.72	0.79	47.89	0.88	0.65	0.87	34.21	0.9	0.33	0.9
IL-17A	0.81	[0.67-0.96]	9.83	0.74	0.82	61.77	0.71	1.68	0.07	57.75	0.46	1.09	0.02
IL-18	0.73	[0.56-0.9]	346.26	0.66	0.77	55.42	0.36	0.97	0.63	20.87	0.73	0.36	0.87
IL-22	0.76	[0.6-0.92]	83.65	0.74	0.75	65.21	0.84	0.87	0.46	26.68	1	0.22	1
IL-36β/IL-1F8	0.77	[0.61-0.93]	11.62	0.74	0.78	38.57	1	0.44	1	32.73	0.96	0.25	1
E-Selectin	0.71	[0.54-0.88]	67632.87	0.64	0.77	18.94	1	0.38	1	6.82	0.95	0.16	1
ICAM-1	0.78	[0.63-0.93]	489976.3	0.74	0.74	93.12	0.06	1.28	0.27	31.26	0.48	0.41	0.75
VCAM-1	0.77	[0.61-0.92]	2992116	0.76	0.69	26.32	0.76	0.76	0.84	12.91	0.93	0.27	0.99
MMP-1	0.78	[0.62-0.94]	1294.94	0.79	0.74	105.17	0.02	2.11	0.04	78.36	0.03	1.29	0.01
MMP-8	0.78	[0.62-0.94]	16814.91	0.78	0.74	40.65	0.9	0.97	0.54	26.08	0.92	0.6	0.33
MMP-9	0.6	[0.41-0.79]	17904.72	0.6	0.67	26.01	0.87	0.41	1	19.39	0.91	0.25	1
C5/C5a	0.6	[0.41-0.8]	30650.78	0.58	0.74	28.39	0.15	0.76	0.85	32.52	0.06	0.58	0.54
Collagen-IVα1	0.73	[0.57-0.9]	1625.95	0.7	0.71	17.81	0.99	0.42	1	10.21	0.99	0.28	1
Fas-Ligand	0.55	[0.35-0.74]	26.12	0.63	0.57	46.23	0.55	0.7	0.89	44.66	0.32	0.38	0.85
Galectin-3	0.7	[0.52-0.88]	2704.71	0.64	0.8	53.72	0.19	1.58	0.09	57.85	0.09	1.45	0.01
G-CSF	0.71	[0.54-0.89]	407.1	0.68	0.74	34.56	0.94	0.46	1	25.12	0.94	0.22	1
I-α-1/COL1A1	0.65	[0.46-0.84]	7385.2	0.78	0.6	42	0.29	0.5	1	33.83	0.34	0.29	0.99
MPO	0.67	[0.48-0.85]	30100.62	0.68	0.74	30.53	0.8	0.46	1	20.06	0.81	0.28	1
Pentraxin-3	0.73	[0.56-0.9]	5726.7	0.69	0.75	12.05	1	0.47	1	9.11	1	0.3	0.97
Resistin	0.77	[0.61-0.93]	34604.01	0.73	0.79	38.53	0.71	0.7	0.93	15	0.79	0.22	1
S100A8	0.72	[0.54-0.9]	547.2	0.74	0.74	21.63	1	1.14	0.16	24.16	0.99	0.85	0.06
S100A9	0.7	[0.51-0.9]	1249.99	0.74	0.82	65.64	0.05	1.81	0.08	47.8	0.08	1.28	0.04
Thrombomodulin	0.95	[0.89-1]	7566.85	0.92	0.89	217.09	0.01	6.36	0.01	111.07	0.01	2.54	0.01
TNFα	0.84	[0.72-0.97]	18.5	0.79	0.77	62.9	0.87	1.69	0.03	54.8	0.68	0.8	0.06
Age	NA	NA	NA	NA	NA	NA	NA	NA	NA	28.14	0.15	0.59	0.51
Sex	NA	NA	NA	NA	NA	NA	NA	NA	NA	-4.59	0.8	0.01	1
SOFA	NA	NA	NA	NA	NA	NA	NA	NA	NA	29.43	0.59	0.64	0.06
SS	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.51	0.87	0.03	0.48
Type	NA	NA	NA	NA	NA	NA	NA	NA	NA	-4.19	0.98	0.02	0.94
WBC	0.58	[0.38-0.78]	16.64	0.58	0.71	NA	NA	NA	NA	2.05	0.96	0.17	1
CRP	0.59	[0.4-0.79]	271.16	0.6	0.66	NA	NA	NA	NA	56.92	0.02	0.86	0.12
Creatinine	0.71	[0.53-0.89]	105.29	0.67	0.77	NA	NA	NA	NA	17.31	0.5	0.24	0.99

Figure 6.5: Thrombomodulin is a plasma protein with biomarker potential for discrimination of NSTI and non-NSTI. Concentrations of the soluble factors in plasma were compared among NSTI patients (n = 251), surgical controls (S. control; n = 20), and non-NSTI controls (n = 20). (A) The median protein levels in each cohort are depicted in the heatmap. All individual values are shown in Figure **??**. The measured proteins are divided by categories: I, chemokines; II, interleukins; III, soluble adhesion molecules; IV, matrix metalloproteases; and V, others. Significant differences between the measured concentrations were tested using Kruskal-Wallis (KW) test followed by Dunn's post hoc test or Mann-Whitney U test (MW). Asterisks indicate the q cutoff obtained in at least 95% of the iterations. *q = 0.05; **q = 0.01; ***q = 0.005. The AUCs from the ROC analyses are given as the mean values of the iterations. The confidence intervals, specificities, and sensitivities of this test are included in table 6.2. (B) The RF result for discriminating NSTI versus non-NSTI is presented as the mean decrease Gini for each variable. The displayed P-values are the result of the model including clinical data (table 6.3). SS, septic shock; type, microbiological classification of NSTI.

6.3.3 Biomarkers discriminating between type I and type II NSTIs

Comparison of the inflammatory response in the 2 types of NSTIs revealed distinct profiles (Figure 6.7A). Type II NSTI patients tended to have higher concentrations of the inflammatory markers, while type I had higher levels of the MMPs. Among the 20 biomarkers with significant differences between the NSTI types, only 6 (i.e., CXCL-10/IP-10, IL-2, IL-10, IL-22, MMP-9, Fas-Ligand) had an AUC greater than 0.7 (Table 6.4), suggesting discriminatory potential. The same set of biomarkers was identified as predictive when the multivariate RF analysis was applied (Figure 6.7B and Table 6.4). Notably, CXCL-10/IP-10 was the biomarker with the highest AUC (0.83; Table 6.4) in the univariate analysis as well as the highest mean decreased Gini in the RF model with a significant P-value (<0.05).

6.3.4 In vitro testing of biomarkers for type differentiation

To further validate the type-specific biomarker panel, we tested to determine whether representative type I and type II NSTI clinical bacterial strains trigger differential inflammatory responses in line with those noted in patient plasma. For this purpose, human peripheral blood mononuclear cells (PBMCs) from healthy donors were stimulated with clinical NSTI bacterial strains. One GAS strain (emm1 type; strain 2006) was selected for the type II infection, while a mix of equal parts of *B. fragilis* and *E. coli* isolated from the same NSTI patient (patient 4011) was used to model a type I NSTI. These species were selected, as they were most frequently cultured in type I patients in the INFECT cohort **(M. B. Madsen, Skrede, et al. 2019)**. The bacterial stimuli included both supernatants containing extracellular factors as well as heat-killed (HK) bacteria for surface-attached factors. Although part of the biomarker panel, MMP-9 was excluded from the in vitro experiment, as PBMCs are not a major cellular source of this factor **(Yabluchanskiy et al. 2013)**. In line with the different plasma concentrations, elevated levels of IL-2, IL-22, CXCL-10/IP-10, and Fas-Ligand were found in type II– versus type I–stimulated cultures (Figure 6.8A–D). In contrast, IL-10 was higher in cells stimulated with HK type I isolates versus the type II GAS isolate (Figure 6.8E); therefore, IL-10 showed the opposite result of that seen with

Table 6.4: Results of the ROC analyses and random forest model discriminating type I and type II NSTI in the discovery cohort. The ROC results are presented as the mean values of the iterations. Thresholds are given as pg/ml. The p-values of the random forest results were calculated using 100 permutations of the original dataset. AUC: Area under the curve, CI: Confidence interval, Sn: Sensitivity, Sp: Specificity, MDA: Mean decrease accuracy, MDG: Mean decrease gini; SOFA: Sequential organ failure assessment score, SS: Septic shock, WBC: white blood cells, CRP: C -reactive protein.

Analyte	Receiver operating characteristic					Random Forest							
						Only proteins				with Clinical variables			
	AUC	[95% CI]	Threshold	Sn	Sp	MDA	p-val	MDG	p-val	MDA	p-val	MDG	p-val
CCL-2/MCP-1	0.59	[0.51-0.66]	645.84	0.6	0.56	34.59	0.04	2.7	0.99	22.74	0.14	2.25	0.91
CCL-4/MIP-1β	0.59	[0.51-0.66]	807.17	0.57	0.59	58.42	0.02	2.66	1	42.21	0.03	2.18	0.9
CCL-5/RANTES	0.53	[0.45-0.6]	7411.44	0.5	0.58	3.88	0.26	2.37	1	1.34	0.36	1.96	1
CXCL-8/IL-8	0.51	[0.43-0.59]	52.1	0.64	0.46	54.46	0.04	2.75	1	52.6	0.02	2.43	0.86
CXCL-10/IP-10	0.83	[0.78-0.89]	216.01	0.79	0.79	369.04	0.01	19.89	0.01	321.92	0.01	15.88	0.01
IL-1α	0.65	[0.58-0.72]	50.2	0.7	0.56	52.35	0.03	2.25	1	47.7	0.02	1.94	1
IL-1β	0.52	[0.44-0.6]	13.79	0.46	0.62	36.04	0.02	2.34	1	26.95	0.03	2.02	0.99
IL-2	0.74	[0.67-0.8]	523.28	0.75	0.67	132.37	0.01	6.88	0.01	115.87	0.01	5.61	0.01
IL-4	0.65	[0.58-0.72]	200.11	0.67	0.6	57.07	0.01	2.54	0.99	44.48	0.03	2.03	0.94
IL-6	0.61	[0.53-0.68]	451.85	0.62	0.56	31.78	0.04	1.95	1	23.63	0.1	1.69	1
IL-10	0.7	[0.63-0.77]	38.56	0.73	0.66	138.72	0.01	5.96	0.02	117.95	0.01	4.63	0.01
IL-12p70	0.66	[0.58-0.73]	165.1	0.75	0.53	79.49	0.01	3.19	0.5	75.22	0.01	2.72	0.25
IL-13	0.61	[0.54-0.69]	1280.83	0.62	0.57	30.65	0.08	1.97	1	21.1	0.06	1.58	1
IL-17A	0.67	[0.6-0.74]	19.83	0.67	0.61	106.05	0.01	3.71	0.27	89.56	0.01	3.08	0.18
IL-18	0.6	[0.53-0.68]	471.26	0.63	0.6	8.5	0.17	2.05	1	-2.71	0.48	1.57	1
IL-22	0.73	[0.66-0.8]	104.19	0.77	0.63	152.12	0.01	6.81	0.01	135.53	0.01	5.46	0.01
IL-36β/IL-1F8	0.65	[0.58-0.73]	14.89	0.61	0.66	71.62	0.01	2.41	1	53.63	0.01	1.86	1
E-Selectin	0.67	[0.6-0.74]	109952.8	0.68	0.61	74.1	0.01	3.42	0.65	58.02	0.01	2.78	0.61
ICAM-1	0.62	[0.55-0.7]	595130.4	0.67	0.59	40.08	0.03	2.33	1	52.16	0.03	2.27	0.99
VCAM-1	0.65	[0.57-0.72]	3490787	0.57	0.67	50.67	0.02	3.68	0.67	59.37	0.02	3.24	0.39
MMP-1	0.56	[0.48-0.64]	1679.64	0.63	0.55	43.03	0.02	3.23	0.97	30.76	0.06	2.09	1
MMP-8	0.52	[0.44-0.59]	27284.41	0.6	0.47	40.52	0.07	3.16	0.88	34.77	0.05	2.59	0.68
MMP-9	0.71	[0.64-0.78]	8858.61	0.77	0.61	139.42	0.01	5.92	0.01	112.09	0.01	4.17	0.02
C5/C5a	0.51	[0.43-0.59]	20808.04	0.47	0.61	29.15	0.08	2.69	1	30.95	0.08	2.14	1
Collagen-IVα1	0.62	[0.54-0.69]	2011.76	0.64	0.61	13.23	0.26	2.83	0.98	-1.3	0.45	2.19	1
Fas-Ligand	0.75	[0.69-0.82]	26.15	0.76	0.68	173.75	0.01	8.16	0.01	160.03	0.01	6.78	0.01
Galectin-3	0.52	[0.44-0.6]	3538.7	0.45	0.65	17.74	0.18	2.37	1	21.67	0.1	2.06	1
G-CSF	0.68	[0.61-0.75]	987.78	0.7	0.6	79.39	0.02	2.98	0.93	72.78	0.01	2.58	0.6
I-α-1/COL1A1	0.66	[0.59-0.73]	7214.79	0.58	0.71	16.78	0.13	3.01	1	21.27	0.1	2.55	0.93
MPO	0.52	[0.44-0.6]	37339.36	0.56	0.51	39.43	0.04	2.63	1	20.85	0.06	1.83	1
Pentraxin-3	0.61	[0.53-0.68]	9159.1	0.59	0.62	42.35	0.06	2.56	1	45.98	0.02	2.34	0.95
Resistin	0.54	[0.46-0.61]	54190.86	0.59	0.51	12.62	0.17	2.5	1	9.84	0.17	1.99	1
S100A8	0.63	[0.56-0.71]	732.57	0.63	0.61	50.82	0.03	2.15	1	44.01	0.04	1.8	1
S100A9	0.59	[0.52-0.67]	3596.5	0.52	0.67	54.82	0.02	3.2	0.94	34.85	0.06	2.18	1
Thrombomodulin	0.52	[0.44-0.6]	11441.51	0.5	0.56	-9.44	0.64	1.89	1	-18.07	0.8	1.44	1
TNFα	0.63	[0.56-0.71]	26.75	0.66	0.6	44.68	0.04	2.11	1	32.18	0.07	1.67	1
Age	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.42	0.42	1.27	1
Sex	NA	NA	NA	NA	NA	NA	NA	NA	NA	-3.1	0.47	0.18	0.99
SOFA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2.87	0.29	1.12	1
SS	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.33	0.26	0.14	0.98
WBC	0.59	[0.51-0.67]	14.08	0.5	0.67	NA	NA	NA	NA	33	0.08	1.97	1
CRP	0.64	[0.57-0.72]	290.74	0.61	0.61	NA	NA	NA	NA	58.37	0.01	3.07	0.4
Creatinine	0.59	[0.51-0.67]	134.04	0.56	0.63	NA	NA	NA	NA	3.52	0.38	1.56	1

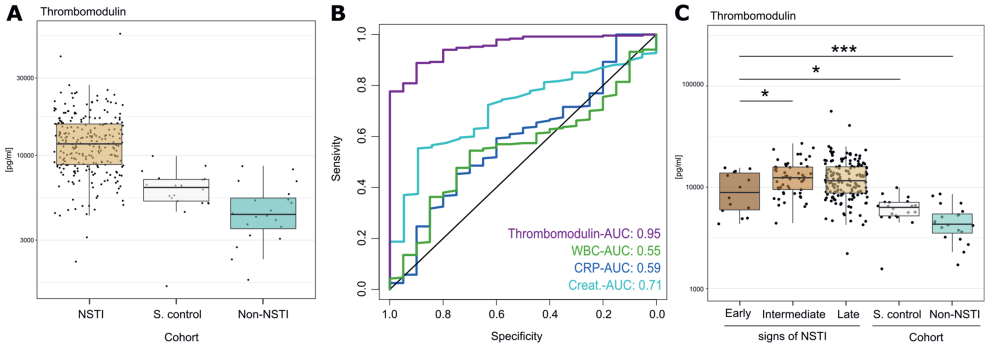


Figure 6.6: Significantly elevated concentration of Thrombomodulin in the NSTI discovery cohort as compared to both the surgical control and non-NSTI control cohorts. (A) Measured levels of Thrombomodulin. (B) ROC curves of Thrombomodulin and selected clinical markers. The legend shows the AUC (Area under the curve) of each curve. (C) NSTI patients were categorized in three groups based on NSTI signs: Early (severe pain, in need of opioids), intermediate (skinbullae or skin bruising) and late (skin purple/black discoloration, skin anesthesia, palpable gas(crepitus) or gas visualised on radiology) signs of NSTI. Statistical testing was performed using Kruskal-Wallis (KW) test followed by Dunn's post hoc test. The resulting p-value of the KW test was 7.62×10^{-15} and the significant differences between patients with early signs and the different groups are shown with stars (*= p-value < 0.05, and ***= p-value<0.005). The boxplots depict the median and the first and third quartiles, whisker extends to the smallest and largest values or 1.5 times the inter-quartile range at most. NSTI: Narcotising soft tissue infection, S. controls: Surgical controls, Non-NSTI: Suspected NSTI but no necrotic tissue found upon surgical exploration, WBC: white blood cells, CRP: C-reactive protein.

the patient data.

6.3.5 Biomarkers discriminating between NSTIs with or without septic shock

Biomarkers associated with severe outcome of NSTIs, such as septic shock, amputation, or death, were also explored within the NSTI cohort. The analyses revealed no significant changes linked to amputation or fatal outcome (Figure 6.9), whereas septic shock was linked to marked differences in inflammatory profile (Figure 5A). Most analytes were significantly higher in plasma of patients with septic shock ($q < 0.05$). However, this was particularly evident in type II cases with or without septic shock, while in type I cases, fewer markers were significantly different (Figure 6.10A). Notably, 3 plasma proteins (i.e., IL-6, granulocyte CSF [G-CSF], and S100A8) were identified as potential biomarkers for septic shock regardless of NSTI type (Figure 6.10B and Tables 6.5 and 6.6).

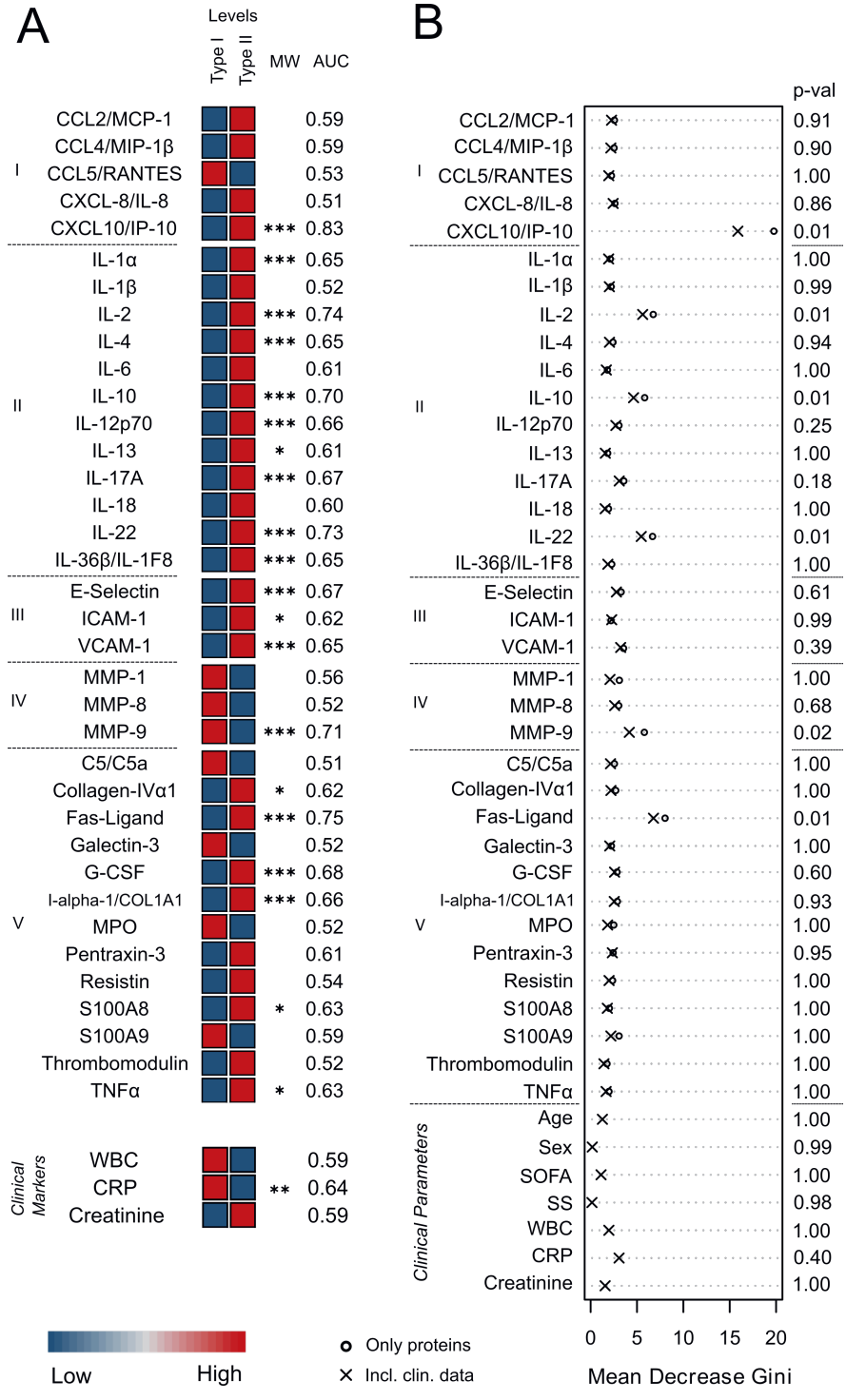

A
Levels
Type I
Type II
MW
AUC
I
CCL2/MCP-1 0.59
CCL4/MIP-1β 0.59
CCL5/RANTES 0.53
CXCL-8/IL-8 0.51
CXCL10/IP-10 *** 0.83
II
IL-1α *** 0.65
IL-1β 0.52
IL-2 *** 0.74
IL-4 *** 0.65
IL-6 0.61
IL-10 *** 0.70
IL-12p70 *** 0.66
IL-13 * 0.61
IL-17A *** 0.67
IL-18 0.60
IL-22 *** 0.73
IL-36β/IL-1F8 *** 0.65
III
E-Selectin *** 0.67
ICAM-1 * 0.62
VCAM-1 *** 0.65
IV
MMP-1 0.56
MMP-8 0.52
MMP-9 *** 0.71
V
C5/C5a 0.51
Collagen-IVα1 * 0.62
Fas-Ligand *** 0.75
Galectin-3 0.52
G-CSF *** 0.68
I-alpha-1/COL1A1 *** 0.66
MPO 0.52
Pentraxin-3 0.61
Resistin 0.54
S100A8 * 0.63
S100A9 0.59
Thrombomodulin 0.52
TNFα * 0.63
Clinical Markers
WBC 0.59
CRP ** 0.64
Creatinine 0.59
Low
High
B
p-val
CCL2/MCP-1 0.91
CCL4/MIP-1β 0.90
CCL5/RANTES 1.00
CXCL-8/IL-8 0.86
CXCL10/IP-10 0.01
IL-1α 1.00
IL-1β 0.99
IL-2 0.01
IL-4 0.94
IL-6 1.00
IL-10 0.01
IL-12p70 0.25
IL-13 1.00
IL-17A 0.18
IL-18 1.00
IL-22 0.01
IL-36β/IL-1F8 1.00
E-Selectin 0.61
ICAM-1 0.99
VCAM-1 0.39
MMP-1 1.00
MMP-8 0.68
MMP-9 0.02
C5/C5a 1.00
Collagen-IVα1 1.00
Fas-Ligand 0.01
Galectin-3 1.00
G-CSF 0.60
I-alpha-1/COL1A1 0.93
MPO 1.00
Pentraxin-3 0.95
Resistin 1.00
S100A8 1.00
S100A9 1.00
Thrombomodulin 1.00
TNFα 1.00
Clinical Parameters
Age 1.00
Sex 0.99
SOFA 1.00
SS 0.98
WBC 1.00
CRP 0.40
Creatinine 1.00
0
5
10
15
20
Mean Decrease Gini
Only proteins
Incl. clin. data

Figure 6.7

Table 6.5: Results from the ROC analyses discriminating septic shock and non-septic shock presence in all cases, type I, and type II NSTI in the discovery cohort. The results are shown as the mean values of the iterations. Thresholds are given as pg/ml. Type I: Polymicrobial culture, Type II: monomicrobial culture, AUC: Area under the curve, CI: Confidence interval, Sn: Sensitivity, Sp: Specificity, WBC: White blood cells, CRP: Creactive protein.

Analyte	All					Type I					Type II				
	AUC	[95% CI]	Threshold	Sn	Sp	AUC	[95% CI]	Threshold	Sn	Sp	AUC	[95% CI]	Threshold	Sn	Sp
CCL-2/MCP-1	0.75	[0.68-0.82]	620.8	0.69	0.76	0.69	[0.58-0.8]	557.19	0.64	0.71	0.79	[0.7-0.88]	625.11	0.77	0.78
CCL-4/MIP-1β	0.72	[0.65-0.79]	793.25	0.75	0.68	0.66	[0.55-0.77]	793.06	0.67	0.72	0.78	[0.68-0.87]	827.33	0.71	0.74
CCL-5/RANTES	0.53	[0.46-0.61]	7656.11	0.58	0.5	0.51	[0.4-0.63]	7155.43	0.56	0.51	0.56	[0.44-0.67]	6179.53	0.55	0.6
CXCL-8/IL-8	0.75	[0.69-0.82]	38.95	0.72	0.71	0.76	[0.67-0.86]	47.59	0.66	0.76	0.75	[0.65-0.85]	36.18	0.74	0.74
CXCL-10/IP-10	0.67	[0.6-0.74]	324.19	0.59	0.67	0.51	[0.39-0.62]	103.24	0.49	0.62	0.7	[0.6-0.8]	1297.83	0.64	0.67
IL-1α	0.75	[0.69-0.82]	49.05	0.67	0.78	0.72	[0.61-0.82]	46.28	0.63	0.76	0.77	[0.68-0.86]	51.51	0.73	0.77
IL-1β	0.69	[0.62-0.76]	15.94	0.63	0.68	0.67	[0.56-0.77]	18.19	0.57	0.74	0.72	[0.62-0.82]	15.16	0.69	0.7
IL-2	0.72	[0.65-0.79]	498.43	0.66	0.69	0.66	[0.55-0.77]	371.54	0.61	0.7	0.75	[0.65-0.84]	663.96	0.73	0.66
IL-4	0.76	[0.69-0.82]	201.45	0.69	0.77	0.69	[0.58-0.8]	178.91	0.68	0.68	0.79	[0.7-0.89]	206.75	0.79	0.76
IL-6	0.78	[0.71-0.84]	419.48	0.71	0.78	0.74	[0.64-0.84]	402.75	0.66	0.81	0.79	[0.71-0.88]	608.89	0.72	0.79
IL-10	0.75	[0.68-0.81]	32.1	0.71	0.7	0.7	[0.6-0.8]	22.01	0.6	0.76	0.77	[0.68-0.86]	60.92	0.67	0.78
IL-12p70	0.72	[0.65-0.79]	154.45	0.65	0.8	0.68	[0.57-0.79]	138.16	0.6	0.77	0.74	[0.64-0.84]	155.7	0.73	0.75
IL-13	0.72	[0.65-0.79]	1276.59	0.67	0.71	0.7	[0.59-0.8]	1274.32	0.61	0.77	0.72	[0.61-0.82]	1306.58	0.68	0.68
IL-17A	0.69	[0.62-0.76]	19.92	0.62	0.69	0.64	[0.53-0.75]	13.14	0.73	0.61	0.72	[0.62-0.83]	23.6	0.69	0.73
IL-18	0.63	[0.55-0.7]	458.2	0.59	0.59	0.6	[0.48-0.71]	457.81	0.51	0.69	0.61	[0.5-0.72]	455.79	0.69	0.51
IL-22	0.72	[0.66-0.79]	98.66	0.7	0.72	0.67	[0.56-0.78]	94.12	0.59	0.75	0.74	[0.64-0.84]	109.44	0.73	0.69
IL-36β/IL-1F8	0.74	[0.68-0.81]	15.91	0.64	0.8	0.67	[0.56-0.78]	13.74	0.66	0.65	0.8	[0.71-0.89]	16.25	0.74	0.8
E-Selectin	0.64	[0.56-0.71]	106330.1	0.61	0.64	0.58	[0.47-0.69]	82682.14	0.58	0.58	0.65	[0.54-0.76]	127423.4	0.66	0.66
ICAM-1	0.63	[0.56-0.71]	590697.9	0.6	0.66	0.59	[0.48-0.71]	545439.1	0.6	0.65	0.66	[0.55-0.76]	663488.8	0.58	0.71
VCAM-1	0.61	[0.54-0.69]	3820864	0.58	0.61	0.49	[0.37-0.6]	2786264	0.58	0.51	0.65	[0.54-0.76]	4147485	0.64	0.65
MMP-1	0.61	[0.54-0.69]	1887.25	0.53	0.64	0.63	[0.51-0.74]	2132.29	0.54	0.66	0.63	[0.52-0.74]	1545.95	0.62	0.62
MMP-8	0.71	[0.64-0.78]	26460.02	0.72	0.65	0.72	[0.61-0.82]	32722.57	0.66	0.73	0.71	[0.61-0.82]	23445.25	0.73	0.69
MMP-9	0.7	[0.63-0.77]	12544.29	0.69	0.67	0.64	[0.54-0.75]	17731.22	0.64	0.64	0.72	[0.62-0.82]	12229.85	0.82	0.6
C5/C5a	0.52	[0.44-0.6]	19591.17	0.51	0.56	0.51	[0.39-0.62]	22388.13	0.6	0.5	0.54	[0.43-0.66]	19283.92	0.51	0.63
Collagen-IVα1	0.75	[0.68-0.81]	1933.49	0.71	0.71	0.72	[0.61-0.82]	1851.14	0.64	0.75	0.76	[0.66-0.85]	2078.03	0.76	0.67
Fas-Ligand	0.66	[0.58-0.73]	26.45	0.6	0.65	0.52	[0.41-0.64]	20.43	0.57	0.58	0.74	[0.64-0.84]	31.58	0.72	0.66
Galectin-3	0.64	[0.57-0.72]	3283.99	0.6	0.64	0.69	[0.59-0.8]	3625.1	0.59	0.75	0.61	[0.5-0.72]	3230.36	0.61	0.63
G-CSF	0.79	[0.73-0.85]	673.45	0.76	0.76	0.77	[0.67-0.86]	501.72	0.7	0.8	0.81	[0.73-0.9]	1487.47	0.75	0.84
I-α-1/COL1A1	0.68	[0.6-0.75]	7805.59	0.66	0.63	0.64	[0.53-0.75]	7428.67	0.56	0.72	0.67	[0.56-0.78]	8715.61	0.66	0.61
MPO	0.68	[0.61-0.75]	40200.88	0.58	0.74	0.63	[0.52-0.74]	43515.31	0.53	0.76	0.73	[0.63-0.83]	36569.84	0.7	0.72
Pentraxin-3	0.74	[0.68-0.81]	7975.61	0.76	0.68	0.69	[0.58-0.79]	7317.16	0.68	0.68	0.77	[0.68-0.87]	9456.21	0.81	0.71
Resistin	0.63	[0.55-0.7]	50956.18	0.65	0.61	0.62	[0.5-0.73]	52193.54	0.63	0.66	0.63	[0.52-0.74]	49776.07	0.69	0.59
S100A8	0.78	[0.71-0.84]	819.42	0.66	0.83	0.75	[0.65-0.84]	661.36	0.69	0.74	0.79	[0.7-0.89]	872.17	0.74	0.84
S100A9	0.55	[0.48-0.63]	3343.11	0.5	0.62	0.55	[0.44-0.67]	4018.29	0.54	0.6	0.58	[0.47-0.69]	2716.59	0.52	0.64
Thrombomodulin	0.62	[0.54-0.69]	12989.07	0.52	0.71	0.54	[0.43-0.66]	13253.96	0.45	0.7	0.68	[0.58-0.79]	12907.4	0.57	0.78
TNFα	0.72	[0.65-0.79]	26.3	0.67	0.71	0.64	[0.53-0.75]	25.64	0.56	0.77	0.77	[0.68-0.86]	30.2	0.71	0.74
WBC	0.6	[0.52-0.68]	14.72	0.55	0.66	0.61	[0.5-0.73]	15.5	0.55	0.69	0.58	[0.47-0.7]	14.02	0.58	0.62
CRP	0.51	[0.43-0.59]	322.79	0.46	0.64	0.57	[0.45-0.68]	383.58	0.47	0.77	0.52	[0.4-0.64]	241.02	0.49	0.64
Creatinine	0.63	[0.56-0.71]	131.1	0.61	0.65	0.55	[0.43-0.66]	115.03	0.55	0.6	0.7	[0.59-0.81]	140.79	0.69	0.68

Table 6.6: Table S4. Results of the random forest models discriminating septic shock and non-septic shock presence in all cases, type I, and type II NSTI in the discovery cohort. The p-values of the random forest results were calculated using 100 permutations of the original dataset. AUC: Area under the curve, p: p-value, CI: Confidence interval, Sn: Sensitivity, Sp: Specificity, MDA: Mean decrease accuracy, MDG: Mean decrease gini, SOFA: Sequential organ failure assessment score, SS: Septic shock, WBC: White blood cells, CRP: C-reactive protein.

Analyte	All								Type I								Type II							
	Only proteins				with Clinical variables				Only proteins				with Clinical variables				Only proteins				with Clinical variables			
	MDA	p	MDG	p	MDA	p	MDG	p	MDA	p	MDG	p	MDA	p	MDG	p	MDA	p	MDG	p	MDA	p	MDG	p
CCL-2/MCP-1	97	0.01	5.2	0.01	85	0.01	3.7	0.03	10	0.31	1.4	0.74	12	0.12	1	0.73	73	0.01	3	0.01	82	0.01	2.7	0.01
CCL-4/MIP-1β	70	0.02	3.6	0.28	79	0.01	3.2	0.06	43	0.05	1.5	0.45	62	0.01	1.7	0.06	52	0.02	1.9	0.1	51	0.01	1.6	0.06
CCL-5/RANTES	-15	0.73	2.3	1	-21	0.89	1.5	1	-25	0.86	0.8	1	-25	0.85	0.6	1	13	0.16	1.3	1	-3	0.33	0.8	1
CXCL-8/IL-8	157	0.01	5.4	0.01	146	0.01	5.5	0.01	101	0.01	3.4	0.02	103	0.01	3	0.01	102	0.01	2.5	0.09	104	0.01	2.5	0.05
CXCL-10/IP-10	51	0.01	2.4	1	7	0.3	1.3	1	5	0.34	1	1	-4	0.39	0.7	1	65	0.01	1.3	0.93	36	0.03	0.8	1
IL-1α	95	0.01	4.2	0.08	79	0.01	3.6	0.04	68	0.02	2.1	0.11	43	0.03	1.6	0.12	50	0.02	1.6	0.33	57	0.02	1.5	0.18
IL-1β	15	0.17	2.6	1	27	0.07	1.8	1	9	0.17	1.2	0.85	9	0.21	0.9	0.96	3	0.39	1.1	1	17	0.19	0.9	0.99
IL-2	44	0.03	2.7	1	45	0.02	2.2	0.88	32	0.09	1.4	0.56	39	0.02	1.4	0.19	30	0.08	1.1	0.99	32	0.06	1	0.89
IL-4	100	0.02	4.8	0.02	97	0.01	4	0.02	40	0.04	1.5	0.26	38	0.04	1.4	0.09	99	0.01	3.1	0.01	98	0.01	2.7	0.01
IL-6	144	0.01	8.8	0.01	123	0.01	6.2	0.01	86	0.01	3.5	0.01	111	0.01	3.3	0.01	98	0.01	3	0.01	77	0.02	1.9	0.05
IL-10	95	0.01	4.1	0.24	64	0.02	2.6	0.45	28	0.11	1.3	0.86	19	0.16	1	0.87	74	0.01	1.8	0.36	58	0.01	1.5	0.24
IL-12p70	61	0.01	4	0.08	33	0.02	2.7	0.15	41	0.02	1.7	0.11	17	0.07	1.2	0.28	16	0.19	1.4	0.79	20	0.14	1	0.67
IL-13	23	0.05	2	1	26	0.05	1.8	0.99	69	0.02	2	0.05	49	0.01	1.7	0.05	11	0.21	0.9	1	9	0.29	0.6	1
IL-17A	45	0.04	2.2	1	51	0.02	1.8	1	16	0.14	1	1	17	0.12	0.8	0.99	22	0.11	1.2	0.99	72	0.01	1.6	0.25
IL-18	49	0.03	2.6	1	53	0.03	1.9	1	-2	0.47	0.9	1	6	0.2	0.8	1	50	0.02	1.6	0.9	48	0.07	1.1	0.97
IL-22	53	0.02	3.8	0.11	38	0.02	2.4	0.51	40	0.05	2	0.12	33	0.04	1.7	0.07	41	0.04	1.2	0.99	34	0.05	0.9	0.99
IL-36β/IL-1F8	95	0.01	4.6	0.01	69	0.01	3.3	0.01	47	0.02	1.3	0.64	25	0.07	1	0.53	93	0.01	3.1	0.01	69	0.02	1.9	0.03
E-Selectin	1	0.41	2.1	1	4	0.37	1.6	1	3	0.28	0.8	1	-9	0.48	0.6	1	7	0.31	1.2	1	14	0.19	1	0.97
ICAM-1	36	0.06	2.3	1	24	0.11	1.5	1	5	0.29	1.1	1	4	0.36	0.8	1	8	0.31	0.8	1	1	0.45	0.7	1
VCAM-1	-7	0.52	2.1	1	-10	0.55	1.4	1	3	0.3	1	1	9	0.27	0.7	1	2	0.37	1.1	1	3	0.3	0.7	1
MMP-1	48	0.03	2.5	1	8	0.25	1.5	1	48	0.05	1.8	0.38	19	0.15	1	0.88	-18	0.83	1	1	-22	0.94	0.8	1
MMP-8	87	0.01	3.3	0.83	79	0.01	2.7	0.49	85	0.01	2.6	0.03	71	0.01	1.9	0.02	27	0.09	1.3	0.98	37	0.04	0.9	1
MMP-9	60	0.01	3.6	0.67	50	0.02	2.2	0.96	8	0.26	1.4	0.83	5	0.3	1	0.96	75	0.03	1.8	0.49	66	0.04	1.4	0.46
C5/C5a	9	0.25	3.2	0.99	-1	0.41	2.1	0.99	-21	0.82	1	1	-18	0.75	0.7	1	30	0.11	1.9	0.55	12	0.23	1.1	0.99
Collagen-IVα1	63	0.01	3.3	0.86	42	0.04	1.9	1	39	0.06	1.5	0.49	35	0.05	1.2	0.54	31	0.04	1.3	1	11	0.26	0.7	1
Fas-Ligand	24	0.04	2.4	1	30	0.1	1.6	1	-1	0.41	1	1	2	0.3	0.8	1	80	0.02	1.7	0.65	73	0.03	1.4	0.43
Galectin-3	43	0.04	2.6	1	63	0.02	2.3	0.94	41	0.07	1.5	0.59	44	0.04	1.5	0.27	0	0.38	0.8	1	27	0.11	0.8	1
G-CSF	196	0.01	11.1	0.01	150	0.01	6.5	0.01	133	0.01	4.2	0.01	112	0.01	2.8	0.01	126	0.01	4.5	0.01	125	0.01	3.7	0.01
I-α-1/COL1A1	21	0.07	2.8	1	34	0.08	2.1	0.98	30	0.1	1.7	0.49	24	0.07	1.2	0.72	-9	0.48	1.1	1	-4	0.44	0.8	1
MPO	92	0.01	3.2	0.96	69	0.01	2.3	0.92	12	0.25	1.3	0.87	22	0.11	1.1	0.66	74	0.01	1.9	0.36	50	0.03	1.1	0.91
Pentraxin-3	117	0.01	6.1	0.01	81	0.02	3.3	0.13	30	0.11	1.5	0.59	9	0.17	0.9	0.97	93	0.02	3.3	0.03	104	0.01	2.8	0.02
Resistin	23	0.1	2.4	1	-1	0.43	1.6	1	19	0.18	1.6	0.59	2	0.34	1.1	0.75	-28	0.93	0.7	1	-14	0.68	0.5	1
S100A8	160	0.01	7.7	0.01	162	0.01	7.2	0.01	115	0.01	3.2	0.01	92	0.01	2.4	0.02	108	0.01	3.9	0.01	123	0.01	3.7	0.01
S100A9	19	0.13	2.3	1	14	0.24	1.8	1	13	0.15	1	1	6	0.29	0.8	1	19	0.12	1.1	1	15	0.13	0.8	1
Thrombomodulin	53	0.02	3.1	1	29	0.02	2.5	0.82	-2	0.48	1.3	0.94	-1	0.33	1.1	0.87	44	0.03	2	0.36	31	0.06	1.3	0.67
TNFα	72	0.02	3.5	0.5	62	0.02	3	0.18	4	0.21	1.2	0.72	10	0.13	1	0.72	62	0.04	2	0.17	66	0.01	2	0.04
Age	NA	NA	NA	NA	-1	0.37	1.7	1	NA	NA	NA	NA	-4	0.5	0.9	0.95	NA	NA	NA	NA	25	0.08	1	0.96
Sex	NA	NA	NA	NA	22	0.03	0.4	0.13	NA	NA	NA	NA	41	0.01	0.6	0.01	NA	NA	NA	NA	0	0.26	0.1	0.96
SOFA	NA	NA	NA	NA	165	0.01	4.2	0.01	NA	NA	NA	NA	75	0.01	1.5	0.04	NA	NA	NA	NA	117	0.01	2	0.02
WBC	NA	NA	NA	NA	15	0.23	1.9	1	NA	NA	NA	NA	-4	0.36	1	0.94	NA	NA	NA	NA	10	0.26	0.8	1
CRP	NA	NA	NA	NA	24	0.08	1.9	1	NA	NA	NA	NA	11	0.19	1	0.92	NA	NA	NA	NA	22	0.13	1	1
Creatinine	NA	NA	NA	NA	23	0.14	2	1	NA	NA	NA	NA	14	0.16	1.1	0.81	NA	NA	NA	NA	21	0.09	0.9	1

Figure 6.7: Biomarker panel for discrimination of type I and type II. Levels of the soluble factors in plasma were compared between type I (n = 117) and type II (n = 134) patients within the NSTI discovery cohort (Table 1). (A) Heatmap depicting the median protein levels in each NSTI type. The measured proteins are divided by categories: I, chemokines; II, interleukins; III, soluble adhesion molecules; IV, matrix metalloproteases; and V, others. Significant differences between the measured concentrations were tested using Mann-Whitney U test. Asterisks indicate the q cutoff obtained in at least 95% of the results. *q = 0.05; **q = 0.01; ***q = 0.005. AUCs from the ROC analyses are shown as the mean values of the iterations. The confidence intervals, specificities, and sensitivities of this test are shown in Table 6.5. (B) The RF result is shown as the mean decrease Gini for each variable. The displayed P-values are the result of the model including clinical data (Table 6.5).

6.3.6 Validation of identified biomarker panels in additional patient cohorts

To test the veracity of the biomarker panels for identification of NSTIs and associated clinical phenotypes (microbiological aetiology and septic shock), a validation cohort was analysed. This cohort comprised 60 additional NSTI patients from the INFECT study (Figure 6.1). To further test the septic shock biomarker panel, 24 patients with sepsis (42% septic shock; no NSTI) of varying aetiology were included (Table 6.7). Due to the exclusive nature of the non-NSTI control group, it was not possible to retrieve similar samples for validation, and instead the sepsis cohort was also used as a comparative cohort to test the predictive value of the NSTI-associated biomarker Thrombomodulin. The discovery and validation NSTI cohorts were well matched with respect to age, sex, and severity of infection. However, the microbiological aetiology differed between cohorts, with GAS being significantly more prevalent in the discovery cohort (Tables 6.7 and 6.8).

All selected biomarkers, including Thrombomodulin, CXCL-10/IP-10, IL-10, MMP-9, G-CSF, S1000A8, IL-6, IL-2, Fas-Ligand, and IL-22, were measured in the validation cohort. However, the results of IL-22 were excluded due to a high number of left-censored data. The measured concentrations of the biomarkers were in the same order of magnitude as in the discovery cohort (Figure 6, A–D). The suggested biomarker for necrosis, Thrombomodulin, showed a high discriminatory power for NSTIs even when compared with the heterogeneous sepsis patient group (Figure 6.11A and Table 6.9).

Among the biomarkers discriminating between type I and type II NSTIs, CXCL-10/IP-10, MMP-9, IL-10, Fas-Ligand, and IL-2, only the first 3 showed significant differences between type I and type II, whereas Fas-Ligand and IL-2 did not (Figure 6.11B). The best performance was noted with CXCL-10/IP-10 (AUC, 0.78; Table 6.9). Since the prevalence of GAS in type II NSTI cases was significantly lower in the validation versus the discovery cohort (38% and 65%, respectively; Tables 6.7 and 6.8), we tested the biomarker panel for comparison of type I versus only GAS type II infections. Notably, the predictive power of CXCL-10/IP-10 reached an impressive AUC of 0.99 (Figure 6.11B and Table 6.9).

The biomarker panel associated with septic shock (i.e., IL-6, G-CSF, and S100A8) in NSTI patients was tested, and the results corroborated the previous findings based on the discovery cohort (Figure 6.11C). The value of these 3 biomarkers was also

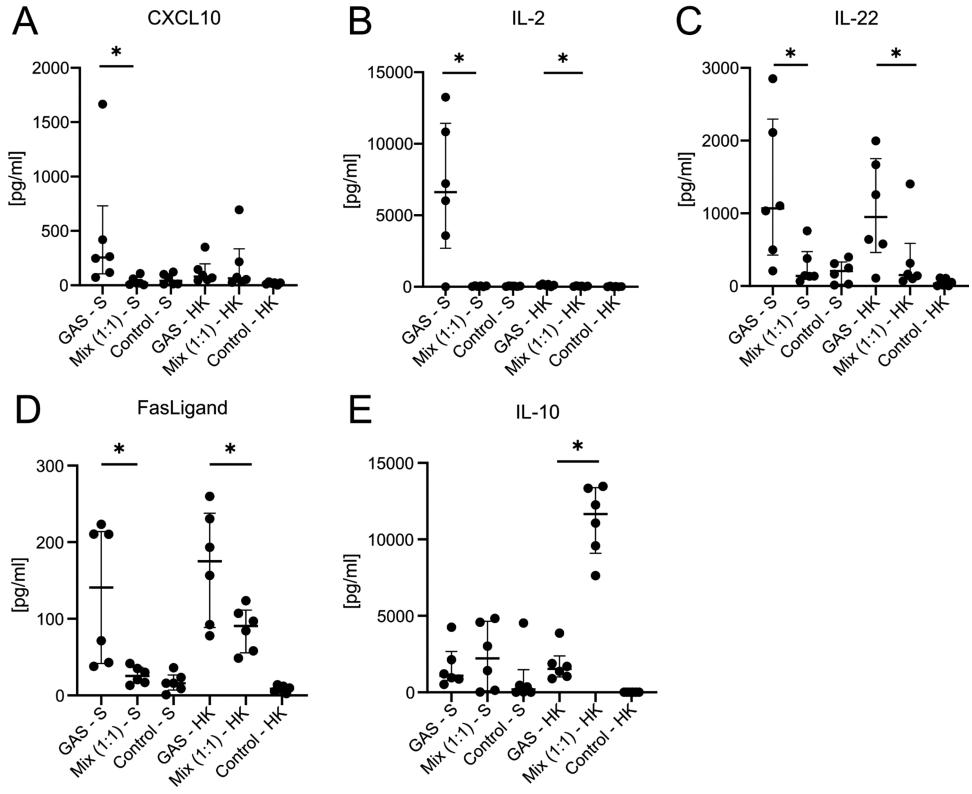


Figure 6.8: Differential production of selected proteins by PBMCs after in vitro stimulation with GAS compared with *B. fragilis* plus *E. coli* (mix). Stimulations were conducted in 6 repeat experiments using PBMCs from different donors stimulated with bacterial supernatant (S) or HK bacteria. (A–E) Scatter plots of each measured analyte in the supernatant after 24 hours of stimulation. The graphs display the individual values and the median with interquartile range. *P < 0.05, Wilcoxon's matched pairs signed rank test.

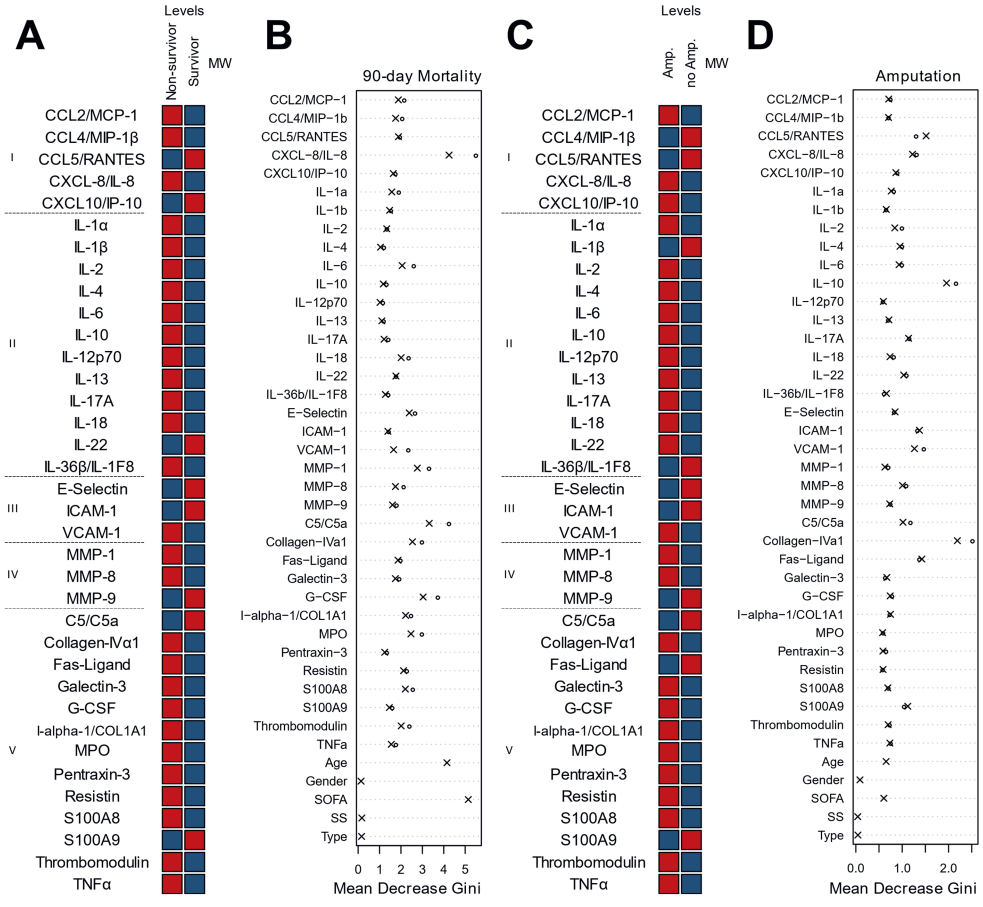


Figure 6.9: None of the measured analytes displayed significant differences for prediction of mortality or amputation in the discovery cohort. (A and C) The median protein levels in each group are depicted in the heatmap. The measured proteins are divided by categories: I-Chemokines, II-Interleukins, III-Soluble adhesion molecules, IV- Matrix metalloproteases, and V-Others. Significant differences between the measured concentrations were tested using Mann-Whitney U test (MW). (B and D) Random forest results shown as the mean decrease Gini for each variable. SOFA: Sequential organ failure assessment score, SS: Septic shock, Amp,: amputation, Type: microbiological classification of NSTI.

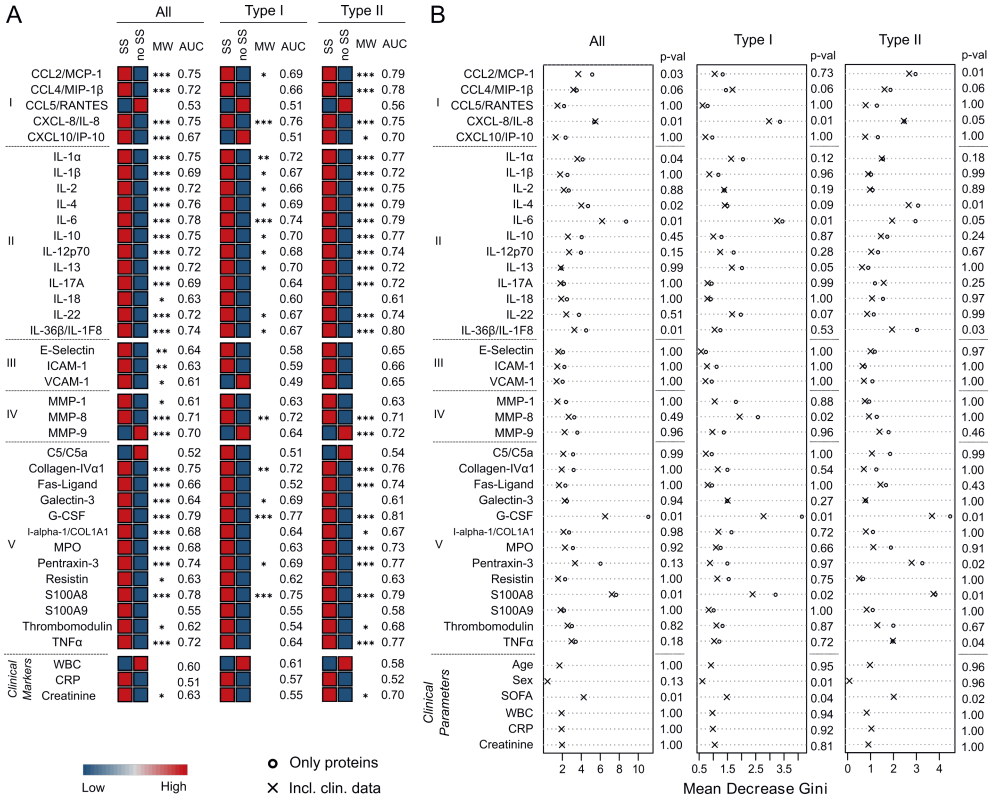


Figure 6.10: Biomarker signatures associated with septic shock differ depending on etiology of NSTI. Levels of the soluble factors in plasma were compared between patients with (n = 134) and without septic shock (n = 117) at admission within the NSTI discovery cohort (Table 1). (A) Heatmaps of the median protein concentrations in each phenotype. The measured proteins are divided by categories: I, chemokines; II, interleukins; III, soluble adhesion molecules; IV, matrix metalloproteases; and V, others. Significant differences between the measured concentrations were tested using Mann-Whitney U test. Asterisks indicate the q value cutoff obtained in at least 95% of the results. *q = 0.05; **q = 0.01; ***q = 0.005. The results from the ROC analysis are shown as the mean AUC values from the iterations. The confidence intervals, specificities, and sensitivities of this test are shown in Table 6.5. (B) RF results are shown as the mean decrease Gini for each variable. Displayed P-values are the results of the models including clinical data (Table 6.4).

Table 6.7: Clinical characteristics of the validation patient cohort and the additional sepsis cohort. The data are shown as mean values and standard deviations, or percentages. NSTI: Narcotising soft tissue infection, type I: Polymicrobial culture, type II: monomicrobial aetiology, NA: not applicable, SAPS II: Simplified acute physiology score II, SOFA: Sequential organ failure assessment score, GAS: Group A Streptococcus, Strep: Streptococcus sp. A Includes only infections in extremities (N=22; Type I=6) B Within 4 weeks before admission for NSTI C Significant differences between cohorts were determined by Mann-Whitney U test or Fisher's exact test.

	NSTI				**p-value** [C]	
	All (n = 60)	**Type I** (n = 39)	**Type II** (n = 21)	**Sepsis** (n = 24)	**NSTI vs Sepsis**	**Type I vs Type II**
Age (years)	58 ± 14	58 ± 11	57 ± 18	69 ± 11	0.001	0.803
Sex (male)	39 (65%)	24 (62%)	15 (71%)	16 (67%)	>0.999	0.573
Septic shock at baseline	29 (48%)	17 (44%)	12 (57%)	11 (46%)	>0.999	0.506
Amputation[A]	3 (14%)	1 (17%)	2 (13%)	NA	NA	>0.999
90-day mortality	8 (13%)	5 (13%)	3 (14%)	4 (17%)	0.735	>0.999
Comorbidities	40 (67%)	26 (67%)	14 (67%)	NA	NA	>0.999
Diabetes (Type I or II)	13 (48%)	11 (61%)	2 (22%)	NA	NA	0.114
Cardiovascular disease	27 (45%)	18 (46%)	9 (43%)	NA	NA	>0.999
Surgery Before NSTI[B]	16 (27%)	14 (36%)	2 (10%)	NA	NA	0.034
SAPS II	44 ± 15 (3% NA)	42 ± 13 (3% NA)	49 ± 17 (5% NA)	NA	NA	0.09
SOFA at admission	8 ± 3 (2% NA)	7 ± 3 (0% NA)	8 ± 4 (5% NA)	4 ± 2 (0% NA)	<0.0001	0.425
Type I	39 (65%)	NA	NA	NA	NA	NA
Microbiological findings						
GAS	8 (13%)	1 (3%)	7 (33%)	2 (8%)	0.717	0.002
Other Strep	3 (5%)	2 (5%)	1 (5%)	0	0.554	>0.999
S. aureus	7 (12%)	2 (5%)	5 (24%)	2 (8%)	>0.999	0.045
Clostridium spp.	1 (2%)	1 (3%)	0 (0%)	0	>0.999	>0.999
Others	41 (68%)	33 (85%)	8 (38%)	20 (83%)	0.188	<0.001

Table 6.8: Statistical comparison of the characteristics of discovery cohort compared to the validation cohort. NSTI: Necrotizing soft tissue infection, Type I: Polymicrobial culture, type II: monomicrobial etiology, NA: not applicable, SAPS II: Simplified acute physiology score II, SOFA: Sequential organ failure assessment score, GAS: Group A Streptococcus, Strep: Streptococcus sp. A Includes only infections in extremities B Within 4 weeks before admission for NSTI C Significant differences between cohorts were determined by Mann- Whitney U test or Fisher's exact test.

	p-value[C]		
	Discovery vs verification		
	All	Type I	Type II
Age (years)	0.445	0.576	0.546
Sex (male)	0.192	0.852	0.103
Septic shock at baseline	0.565	0.853	>0.999
Amputation[A]	0.435	>0.999	>0.999
90-day mortality	0.353	0.252	>0.999
Comorbidities	0.434	0.211	>0.999
Diabetes (Type I or II)	0.866	0.437	>0.999
Cardiovascular disease	0.56	0.581	0.813
Surgery Before NSTI[B]	0.035	0.142	0.667
SAPS II	0.849	0.465	0.366
SOFA at admission	0.141	0.534	0.412
Type I	0.014	NA	NA
Microbiological findings			
GAS	0.0001	0.294	0.007
Other Strep	0.035	0.067	0.314
S. aureus	0.278	0.731	0.012
Clostridium spp.	0.475	0.681	>0.999
Others	<0.0001	0.006	0.003

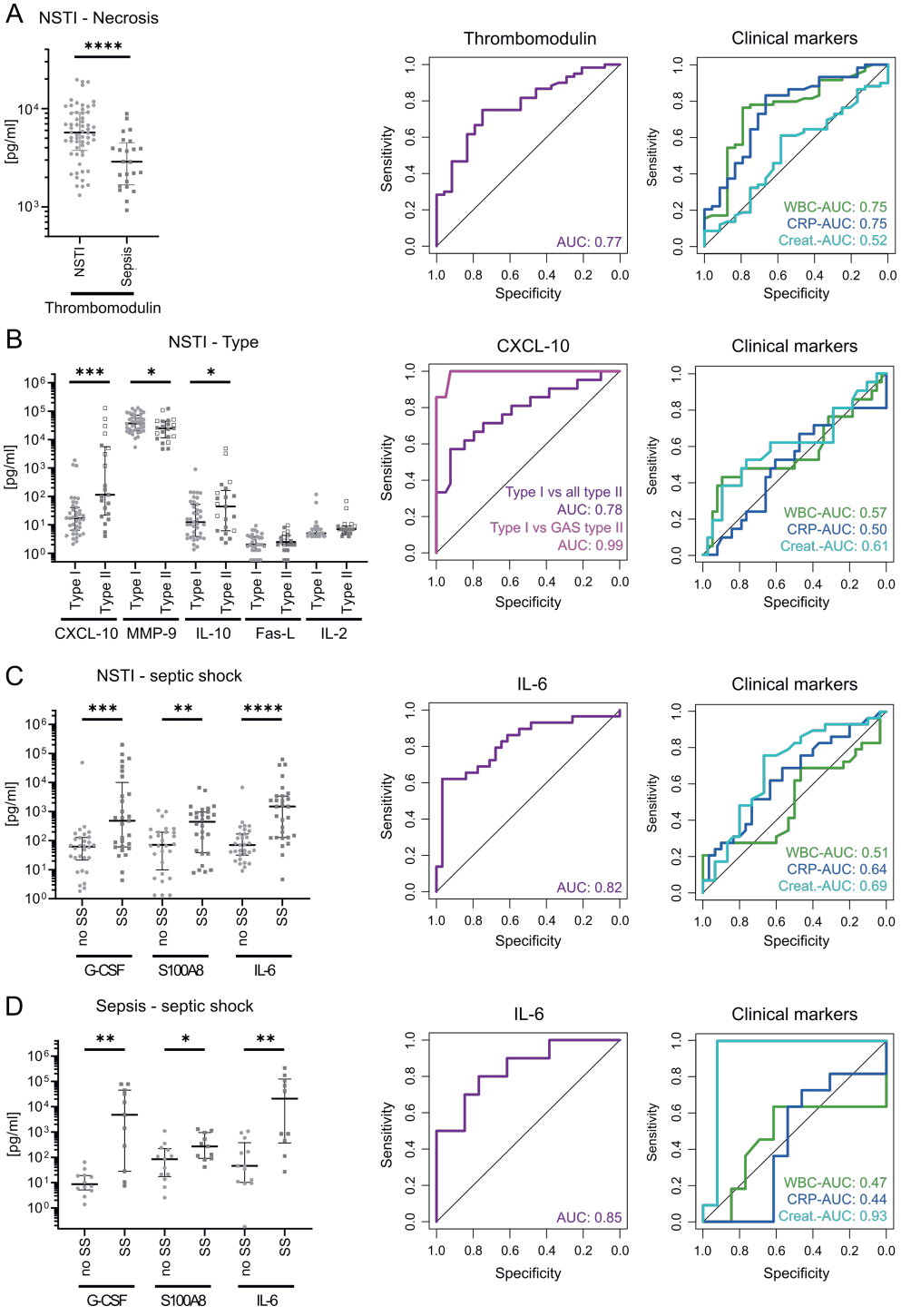

A
NSTI - Necrosis
****
[pg/ml]
NSTI
Sepsis
Thrombomodulin
Thrombomodulin
Sensitivity
Specificity
AUC: 0.77
Clinical markers
WBC-AUC: 0.75
CRP-AUC: 0.75
Creat.-AUC: 0.52
B
NSTI - Type
***
*
*
Type I
Type II
CXCL-10
MMP-9
IL-10
Fas-L
IL-2
CXCL-10
Type I vs all type II
AUC: 0.78
Type I vs GAS type II
AUC: 0.99
Clinical markers
WBC-AUC: 0.57
CRP-AUC: 0.50
Creat.-AUC: 0.61
C
NSTI - septic shock
***
**
****
no SS
SS
G-CSF
S100A8
IL-6
IL-6
AUC: 0.82
Clinical markers
WBC-AUC: 0.51
CRP-AUC: 0.64
Creat.-AUC: 0.69
D
Sepsis - septic shock
**
*
**
IL-6
AUC: 0.85
Clinical markers
WBC-AUC: 0.47
CRP-AUC: 0.44
Creat.-AUC: 0.93

Figure 6.11

Table 6.9: Results of the ROC analyses applied in the validation cohort. Thresholds are given in pg/ml. NSTI: Necrotizing soft tissue infection, type I: Polymicrobial culture, type II: monomicrobial culture, GAS: Group A Streptococcus, SS: Septic shock, CI: Confidence interval, WBC: White blood cells, CRP: C-reactive protein

		Comparison	Analyte	AUC	[95% CI]	Threshold	Sensitivity	Specificity
NSTI		**Sepsis**	Thrombomodulin	0.77	[0.67-0.88]	3972.01	0.75	0.75
		vs NSTI	WBC	0.75	[0.63-0.87]	13.3	0.76	0.79
			CRP	0.75	[0.63-0.87]	153	0.83	0.67
			Creatinine	0.52	[0.39-0.66]	114.5	0.61	0.58
Type	**NSTI**	**Type I**	CXCL-10/IP-10	0.78	[0.65-0.91]	36.43	0.71	0.74
	cohort	**vs**	MMP-9	0.66	[0.51-0.81]	34902.5	0.71	0.54
		Type II	IL-10	0.63	[0.47-0.79]	43.5	0.52	0.74
			Fas-Ligand	0.55	[0.39-0.72]	1.88	0.79	0.42
			IL-2	0.64	[0.49-0.78]	6.02	0.71	0.53
			WBC	0.57	[0.4-0.74]	14.3	0.48	0.79
			CRP	0.5	[0.34-0.66]	242.5	0.52	0.61
			Creatinine	0.61	[0.44-0.77]	106	0.62	0.63
	NSTI	**Type I**	CXCL-10/IP-10	0.99	[0.97-1]	642.77	1	0.92
	cohort	**vs**	MMP-9	0.6	[0.33-0.87]	16960	0.43	0.9
		GAS	IL-10	0.79	[0.6-0.99]	53.41	0.71	0.77
			Fas-Ligand	0.74	[0.56-0.92]	2.61	0.86	0.58
			IL-2	0.82	[0.65-0.99]	8.01	0.71	0.82
			WBC	0.58	[0.28-0.87]	21.2	0.57	0.66
			CRP	0.51	[0.25-0.77]	286	0.57	0.53
			Creatinine	0.7	[0.45-0.95]	150	0.71	0.76
Septic	**NSTI**	**no SS**	G-CSF	0.76	[0.64-0.88]	334.3	0.55	0.97
Shock	**cohort**	**vs SS**	S100A8	0.72	[0.58-0.85]	289.47	0.57	0.89
			IL-6	0.82	[0.71-0.93]	467.5	0.62	0.97
			WBC	0.51	[0.36-0.67]	20.8	0.69	0.47
			CRP	0.64	[0.5-0.78]	265.5	0.62	0.63
			Creatinine	0.69	[0.56-0.83]	97.5	0.76	0.67
	Sepsis	**no SS**	G-CSF	0.83	[0.65-1]	23.66	0.73	0.83
	cohort	**vs SS**	S100A8	0.76	[0.57-0.96]	112.96	0.73	0.69
			IL-6	0.85	[0.68-1]	308.81	0.8	0.77
			WBC	0.47	[0.21-0.74]	10.65	0.64	0.62
			CRP	0.44	[0.19-0.69]	105	0.64	0.54
			Creatinine	0.93	[0.79-1]	117.5	1	0.92

Figure 6.11: Predictive power of plasma biomarkers assessed in the validation cohort. Selected biomarkers were tested for their potential to detect (A) NSTI (necrosis), (B) NSTI type, and (C and D) septic shock. Scatter plots display the individual values and the median with interquartile range. The discovery cohort consist of 60 NSTI patients, of which 39 were type I and 29 developed septic shock (Table 6.7). In panel B, empty squares indicate type II NSTI caused by GAS (n = 7). The control group of 24 sepsis patients included 11 patients with septic shock. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001; Mann-Whitney U test. ROC plots display results of the indicated biomarkers or clinical markers.

tested in the sepsis cohort, revealing a similar discriminatory power (Figure 6.11D and Table 6.9). Among these biomarkers, IL-6 showed the best performance, with AUCs of 0.82 and 0.85 in the NSTI and sepsis cohorts, respectively. In line with creatinine being a definition marker for acute kidney injury and sepsis-associated organ failure, creatinine in the sepsis cohort showed an AUC of 0.93 (Figure 6.11D). Finally, multivariate logistic regression revealed that all biomarkers retained their discriminatory power for the specific NSTI clinical phenotypes even when sex, age, and SOFA scores were considered (Table 6.10).

6.3.7 Network analysis

The specific biomarker panels identified in type I versus type II NSTIs as well as the septic shock profiles displayed by either type implied differential mechanisms underlying the skewed inflammatory responses. To gain further insights into this and to identify key response nodes, network connectivity analysis was applied to assess interactions among the markers in the discovery cohort data set (Figure 6.12 and Figure 6.13). In general, we observed a more densely connected network in NSTIs in comparison with both control groups (Figure 6.13). Whereas most analytes were disconnected in the controls, some analytes, such as Pentraxin-3 and CXCL-8/IL-8, gained many interacting partners in NSTI cases. The results revealed that the differences in connectivity were related not only to the presence of unique connections, but also to the strength of the connections. Furthermore, the analyses revealed striking differences in connectivity patterns even between the different NSTI clinical phenotypes (Figure 6.12). The most pronounced differential connectivities between type I and type II NSTIs were noted for IL-6, IL-1α, and CCL-4/MIP-1β, all of which were stronger in type II. The analytes with a high number of connections (hubs) that displayed significant differential connectivity between septic shock and non septic shock were different for the 2 types of NSTIs. In type I cases, connections among the interleukins IL-1α, IL-4, and IL-17A were the most relevant (q < 0.002), while type II NSTIs displayed the most changes in other analytes, such as Galectin-3, I-α-1/COL1A1, and Thrombomodulin (q < 0.001) (Table 6.11).

6.4 Discussion

In this study, we identify a set of plasma biomarkers that discriminate between distinct clinical NSTI phenotypes. Robust profiles were defined for NSTI versus non-NSTI controls and type I and type II NSTIs as well as septic shock development. A

Table 6.10: Uni- and multivariate logistic regression analyses performed in the validation cohort. NSTI: Necrotizing soft tissue infection, Type I: Polymicrobial culture, Type I: monomicrobial culture, GAS: Group A Streptococcus, SS: Septic shock, CI: Confidence interval, WBC: White blood cells, CRP: C-reactive protein. *Logistic regression including age, sex and SOFA

	Samples		Analyte	Univariate			Multivariate*		
				OR	[95% CI]	p-value	OR	[95% CI]	p-value
NSTI		**Sepsis**	Thrombomodulin	2.79	[1.63-5.18]	<0.001	3.01	[1.42-7.52]	0.003
		vs NSTI	WBC	2.29	[1.41-4.02]	<0.001	2.62	[1.34-5.93]	0
			CRP	2.31	[1.48-3.85]	<0.001	1.85	[1.03-3.5]	0.039
			Creatinine	0.95	[0.54-1.71]	0.86	0.58	[0.23-1.33]	0.2
Type	**NSTI**	**Type I**	CXCL-10/IP-10	1.37	[1.15-1.7]	<0.001	1.39	[1.16-1.76]	<0.001
	cohort	**vs**	MMP-9	0.6	[0.36-0.95]	0.03	0.49	[0.27-0.82]	0.01
		Type II	IL-10	1.23	[1-1.54]	0.045	1.31	[1.04-1.71]	0.023
			Fas-Ligand	1.15	[0.69-1.94]	0.59	1.06	[0.63-1.8]	0.84
			IL-2	1.28	[0.78-2.22]	0.321	1.22	[0.72-2.17]	0.45
			WBC	0.66	[0.35-1.2]	0.17	0.63	[0.32-1.15]	0.14
			CRP	1.06	[0.6-1.92]	0.849	1.05	[0.59-1.91]	0.87
			Creatinine	1.65	[0.9-3.19]	0.11	1.64	[0.86-3.32]	0.14
	NSTI	**Type I**	CXCL-10/IP-10	2.12	[1.43-21.4]	<0.001	1.68	[1.3-5.23]	<0.001
	cohort	**vs**	MMP-9	0.74	[0.35-1.49]	0.39	0.69	[0.31-1.4]	0.31
		GAS	IL-10	1.52	[1.14-2.25]	0.003	1.92	[1.23-3.53]	0.003
			Fas-Ligand	2.16	[0.97-6.26]	0.06	2.64	[1.03-9.94]	0.04
			IL-2	1.78	[1.03-3.33]	0.041	1.96	[0.96-4.73]	0.064
			WBC	0.89	[0.39-2.58]	0.8	0.93	[0.42-2.44]	0.87
			CRP	0.96	[0.45-2.38]	0.928	0.88	[0.39-2.16]	0.768
			Creatinine	2.25	[0.93-6]	0.072	2	[0.73-6.48]	0.181
Septic	**NSTI**	**no SS**	G-CSF	1.33	[1.13-1.67]	<0.001	1.37	[1.14-1.76]	<0.001
Shock	**cohort**	**vs SS**	S100A8	1.3	[1.07-1.61]	0.006	1.25	[1.02-1.59]	0.028
			IL-6	1.61	[1.28-2.18]	<0.001	1.51	[1.19-2.03]	<0.001
			WBC	0.73	[0.38-1.3]	0.286	0.53	[0.18-1.17]	0.125
			CRP	1.68	[0.96-3.2]	0.072	1.48	[0.76-3.06]	0.254
			Creatinine	2.06	[1.1-4.26]	0.023	1.47	[0.71-3.17]	0.304
	Sepsis	**no SS**	G-CSF	1.39	[1.11-2.37]	0.002	1.36	[1.03-2.52]	0.028
	cohort	**vs SS**	S100A8	1.54	[1.05-2.62]	0.027	1.27	[0.71-2.64]	0.431
			IL-6	1.41	[1.12-2.15]	0.001	1.3	[1-1.98]	0.046
			WBC	0.72	[0.33-1.35]	0.308	0.65	[0.23-1.43]	0.286
			CRP	0.79	[0.41-1.43]	0.429	0.54	[0.13-1.48]	0.244
			Creatinine	17.7	[3.03-331.92]	<0.001	21.43	[1.9-6.5E12]	0.006

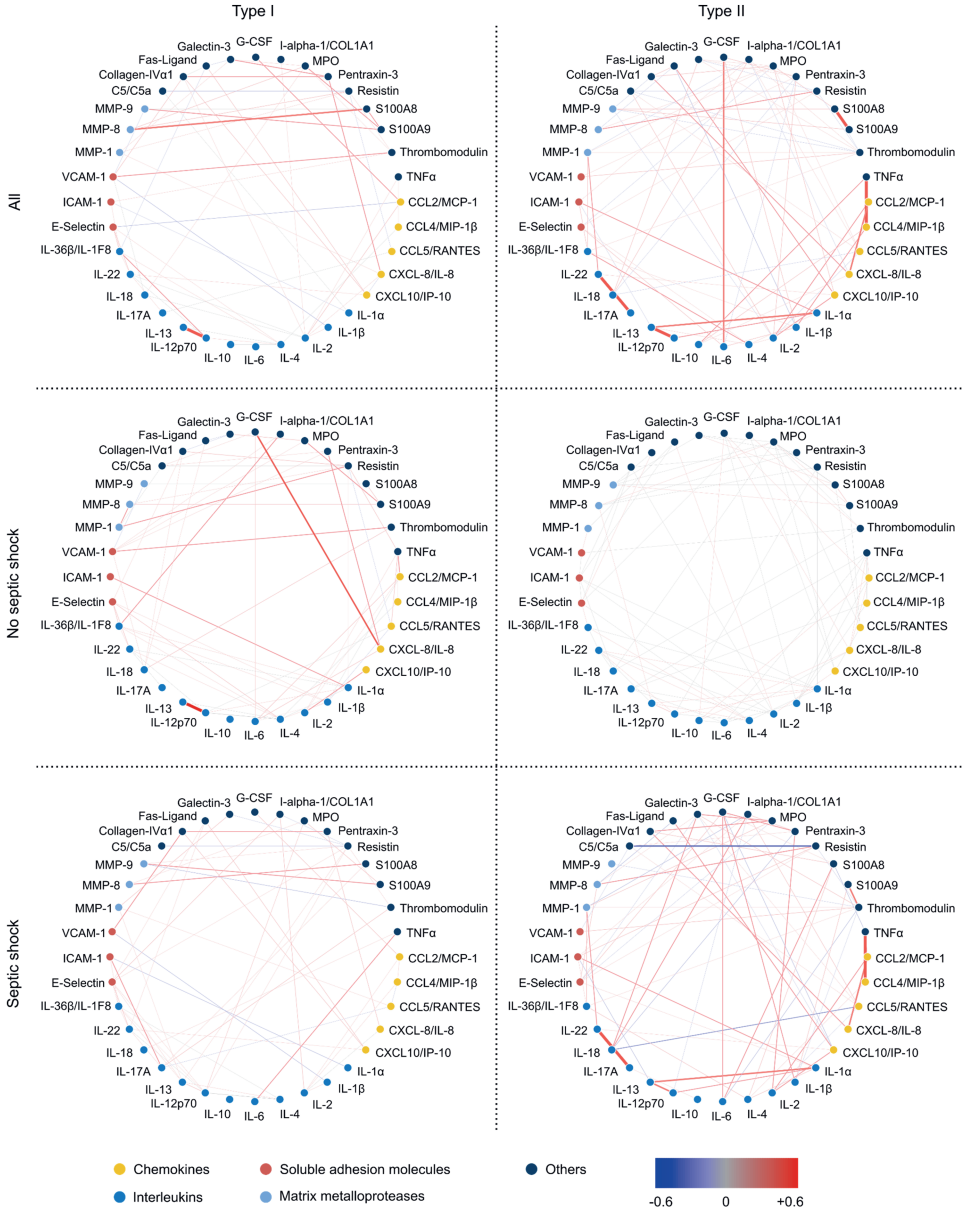


Figure 6.12: Type I and type II NSTIs display contrasting association networks. The colours of the circles indicate the categories of the analytes. The strength of the partial correlation between analytes is indicated by the colour and the weight of the connection.

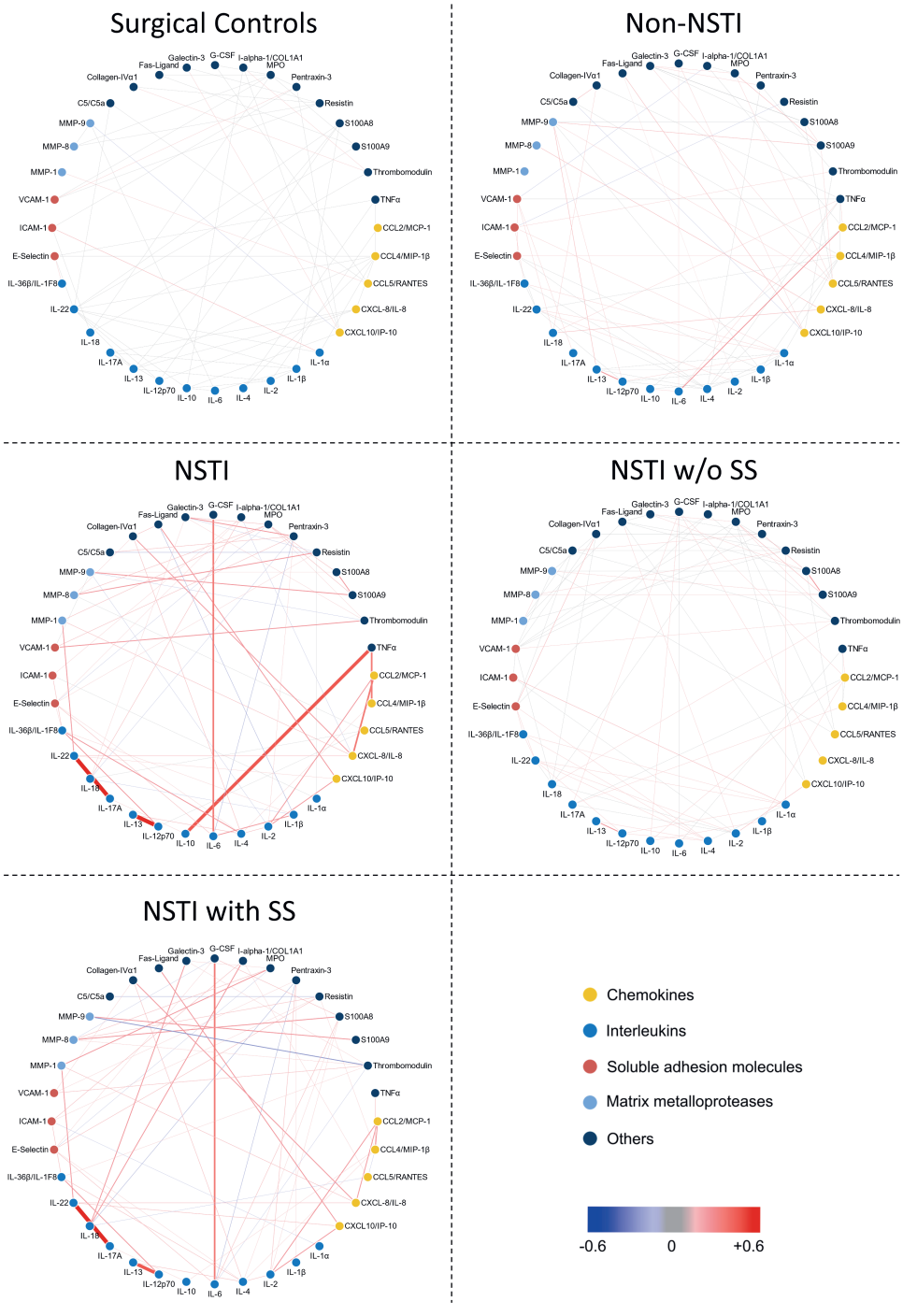

Surgical Controls
Non-NSTI
NSTI
NSTI w/o SS
NSTI with SS
Galectin-3
G-CSF
I-alpha-1/COL1A1
Fas-Ligand
MPO
Collagen-IVα1
Pentraxin-3
C5/C5a
Resistin
MMP-9
S100A8
MMP-8
S100A9
MMP-1
Thrombomodulin
VCAM-1
TNFα
ICAM-1
CCL2/MCP-1
E-Selectin
CCL4/MIP-1β
IL-36β/IL-1F8
CCL5/RANTES
IL-22
CXCL-8/IL-8
IL-18
CXCL10/IP-10
IL-17A
IL-1α
IL-13
IL-1β
IL-12p70
IL-2
IL-10
IL-6
IL-4
Chemokines
Interleukins
Soluble adhesion molecules
Matrix metalloproteases
Others
-0.6
0
+0.6

Figure 6.13

Table 6.11: Connectivity values for each measured analyte in the discovery cohort. Samples were aggregated by type. The differential connectivity and q-values were calculated for the comparison of septic shock development within each NSTI type. DC: Differential connectivity, q: q-values. Categories: I-Chemokines, II-Interleukins, III-Soluble adhesion molecules, IV-Matrix metalloproteases, and V-Others. Type I: Polymicrobial culture, SS: Septic shock.

	Analyte	Type I				Type II			
		Connectivity		DC	q	Connectivity		DC	q
		no SS	SS			no SS	SS		
I	CCL-2/MCP-1	0.74	0.49	0.26	0.002	0.5	0.77	0.27	0.001
	CCL-4/MIP-1β	0.21	0.34	0.13	0.003	0.28	0.55	0.26	0.001
	CCL-5/RANTES	0.49	0.33	0.16	0.01	0.37	0.77	0.39	0.001
	CXCL-8/IL-8	0.7	0.3	0.41	0.002	0.48	0.54	0.06	0.041
	CXCL-10/IP-10	0.26	0.47	0.21	0.002	0.32	0.87	0.55	0.001
II	IL-1α	0.81	0.17	0.64	0.002	0.69	1.07	0.38	0.001
	IL-1β	0.7	0.29	0.4	0.002	0.46	0.51	0.05	0.097
	IL-2	0.62	0.66	0.03	0.313	0.5	0.87	0.37	0.001
	IL-4	0.9	0.29	0.61	0.002	0.61	0.63	0.02	0.505
	IL-6	0.67	0.27	0.39	0.002	0.31	0.68	0.36	0.001
	IL-10	0	0	0	0.833	0.37	0.11	0.26	0.001
	IL-12p70	0.81	0.98	0.17	0.003	0.59	0.68	0.08	0.034
	IL-13	0.49	0.75	0.25	0.002	0.59	0.86	0.27	0.001
	IL-17A	0.09	0.73	0.64	0.002	0.3	0.59	0.28	0.001
	IL-18	0.52	0.13	0.39	0.002	0.3	0.9	0.6	0.001
	IL-22	0.69	0.72	0.02	0.454	0.64	0.85	0.21	0.001
	IL-36β/IL-1F8	0.96	0.85	0.12	0.025	0.42	0.65	0.23	0.001
III	E-Selectin	0.74	0.88	0.14	0.018	0.55	0.87	0.32	0.001
	ICAM-1	0.27	0.55	0.28	0.002	0.53	0.84	0.31	0.001
	VCAM-1	0.84	0.36	0.48	0.002	0.29	0.29	0	0.767
IV	MMP-1	0.54	0.37	0.17	0.003	0.18	0.89	0.71	0.001
	MMP-8	0.55	0.37	0.18	0.002	0.4	0.69	0.28	0.001
	MMP-9	0.2	0.58	0.38	0.002	0.46	0.14	0.31	0.001
V	C5/C5a	0.47	0.13	0.34	0.002	0.31	0.64	0.33	0.001
	Collagen-IVα1	0.51	0.64	0.13	0.01	0.21	0.59	0.38	0.001
	Fas-Ligand	0.21	0.44	0.23	0.002	0.24	0.38	0.14	0.011
	Galectin-3	0.4	0.28	0.13	0.009	0.34	1.19	0.84	0.001
	G-CSF	0.68	0.2	0.48	0.002	0.41	1.15	0.74	0.001
	I-α-1/COL1A1	0.4	0.44	0.04	0.205	0.28	1.03	0.74	0.001
	MPO	0.53	0.42	0.1	0.049	0.83	0.71	0.12	0.006
	Pentraxin-3	0.45	0.48	0.03	0.293	0.27	0.93	0.66	0.001
	Resistin	0.97	1.01	0.04	0.313	0.26	0.89	0.63	0.001
	S100A8	0.29	0.33	0.04	0.205	0.59	0.84	0.25	0.001
	S100A9	0.59	0.42	0.17	0.002	0.53	0.49	0.05	0.075
	Thrombomodulin	0.42	0.46	0.04	0.313	0.49	1.46	0.98	0.001
	TNFα	0.82	0.4	0.42	0.002	0.57	0.74	0.18	0.001

Figure 6.13: Association networks of measured analytes in the surgical control, Non-NSTI control, NSTI, NSTI patients without septic shock, and NSTI patients with septic shock from the discovery cohorts. The non- NSTI patients are suspected NSTI cases but that had no necrotic tissue upon surgical exploration. The strength of the partial correlation between analytes is shown by the colour and the weight of the connection.

key strength of the study is that it is based on the prospective multicenter NSTI patient cohort (the INFECT cohort), which is the largest available NSTI cohort, and it also includes an extensive biobank collected using harmonized standard operating procedures **(M. Madsen et al. 2018)**. To identify analytes with the highest predictive power to discriminate between different clinical phenotypes, a set of stringent statistical analyses was applied to the data set, including uni- and multivariate analyses with embedded resampling to account for unequal patient numbers in specific patient groups. The multivariate analyses included clinical parameters of age, sex, and SOFA score to assess their contributions to the identification of the different clinical phenotypes. The finding of unique predictive biomarker panels related to specific clinical phenotypes suggests differential underlying pathophysiological mechanisms. This was further strengthened by the connectivity analyses demonstrating differential marker-marker interactions as well as different key hubs (i.e., densely connected) in the specific clinical phenotype–linked networks.

The development of rapid diagnostic tools for NSTI, such as levels of disease-associated biomarkers, to support clinical decisions could increase the accuracy of early diagnosis, leading to swifter surgical exploration and treatment only when clinically indicated. However, to date, there are only a few studies of molecular biomarkers in NSTIs **(Hansen, Rasmussen, Svensson, et al. 2017; Lungstras-Bufler et al. 2004; Hansen, Rasmussen, Garred, Bidstrup, et al. 2016; Hansen, Rasmussen, Garred, Pilely, et al. 2018; Polzik et al. 2019; Kristensen et al. 2020)**, and these are limited to analyses of only a few markers. The comprehensive multiplex analysis of 36 analytes conducted here revealed, as expected, a greater systemic inflammatory response in NSTI patients than in non infected patients (surgical controls). Less drastic changes were observed when NSTI patients were compared with the infected non-NSTI controls. This is in line with the non-NSTI cases having a severe soft-tissue infection, to the extent that they were initially suspected NSTIs, but in which no necrosis was found upon surgical exploration, and hence, greater similarity in the host response is reasonable. Notably, Thrombomodulin emerged as a robust candidate for the discrimination of NSTI from non-NSTI, indicating its potential as a biomarker for soft-tissue necrosis. Although there are no published reports exploring Thrombomodulin in soft-tissue infections, it has been linked to necrotising pancreatitis **(X.-L. Lu et al. 2007)**. Further studies are needed to dissect the underlying mechanism leading to elevation of Thrombomodulin and its role in NSTI and necrosis. Our data support that Thrombomodulin is a biomarker of relatively early disease, as it was noted in patients showing only early signs of NSTI. However, these data need to be interpreted with caution, since classifications of early and late signs are based on patient chart notes and, because of this, potential bias cannot be excluded. Therefore, further studies are warranted, and it would be of value to assess samples already collected in the ambulance or the emergency department.

There were no differences between Thrombomodulin levels in type I and type II, which is in agreement with previous reports demonstrating high levels of soluble Thrombomodulin in bacterial infections regardless of the causative microorganism **(X. Guo et al. 2019; Kinasewitz et al. 2004)**. Moreover, elevated concentrations of Thrombomodulin in blood have been reported in patients with sepsis **(Kinasewitz et al. 2004; Iba, Yagi, et al. 1995; J.-J. Lin et al. 2017; Mihajlovic et al. 2015)**. Such elevated levels were also detected in the sepsis cohort we included during the validation stage. Notably, Thrombomodulin retained its discriminatory power for NSTIs. Finally, Thrombomodulin has also been proposed as a biomarker for the prediction of mortality in patients with sepsis **(Kinasewitz et al. 2004; J.-J. Lin et al. 2017)** and septic shock **(Fang et al. 2018)**. Although an association with mortality in NSTI was not noted in our study, a weak association with septic shock was identified.

Identification of biomarkers associated with septic shock in NSTI patients was a key focus of this study, as early identification of this complication is critical for optimal tailored patient management. The plasma inflammatory response profile indicated a septic shock signature that was dependent on the NSTI type. Three septic shock–associated markers, i.e., IL-6, G-CSF, and S100A8, were shared for both types. In the validation stage, we confirmed the discriminatory power of all 3 biomarkers for septic shock. Hence, this confirms their biomarker potential in NSTIs and likely also in other severe infectious diseases, such as sepsis. Additionally, we also explored biomarker signatures for major outcomes, such as death and amputation. We failed to identify a significant biomarker signature related to these outcomes, which may be due to our highly stringent analyses. It should also be noted that amputation as readout is associated with many confounders, such as praxis at the clinical site.

Early targeted antibiotic treatment of NSTIs is critical for the successful management of patients, and therefore, biomarkers for the discrimination of types I and II NSTIs could serve to accelerate the decision-making process in the clinics. In this study, CXCL-10/IP-10, IL-2, IL-10, IL-22, MMP-9, and Fas-Ligand were identified as discriminatory biomarkers for type I and type II infections. Out of these 6 markers, MMP-9 was the only marker with higher concentrations in type I versus type II, whereas the rest were higher in type II. We sought to validate the predictive biomarker sets identified in the discovery cohort by in vitro stimulation experiments to model the type I and type II infections. The measurement of the selected analytes in media from the validation stimulations experiments showed higher levels of CXCL-10/IP-10, IL-2, IL-22, and Fas-Ligand associated with type II, as compared with type I, bacterial stimulation. However, IL-10 responses in in vitro stimulations did not match the variation noted in patient plasma. This discordant result is likely due to the limitations in the in vitro assay failing to mimic the complex in vivo setting. Nonetheless, 4 out of 5 tested biomarkers corroborated the patient data and substantiated the association of specific biomarkers to the type of infection.

Among all tested biomarkers, CXCL-10/IP-10 displayed the strongest power to discriminate type II from type I NSTIs. Although this association was noted in both the discovery and the validation cohort, it was substantially more impressive in the discovery cohort. As there was a difference in the frequency of GAS type II cases between the 2 cohorts, a sub-analysis including only type II GAS cases was performed and revealed an almost perfect differentiation from type I cases. Hence, the relevance of this biomarker is likely connected to GAS rather than to all type II infections. In

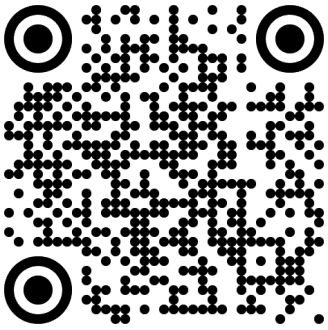
line with this, our recent study using dual RNA-Seq analyses of processed tissue biopsies from NSTI patients revealed a higher expression of CXCL9, CXCL-10/IP-10, and CXCL11 in NSTI GAS type II infections versus type I **(Thänert et al. 2019)**. Moreover, IL-10 and IL-2 have been previously linked to severe and nonsevere GAS infections. IL-10 has been reported to be elevated during GAS infections and significantly higher in invasive versus noninvasive infections **(S.-M. Wang et al. 2008; Osowicki et al. 2021)**. The frequency of IL-2 –producing cells in circulation increased in patients with severe invasive GAS infections **(Norrby-Teglund, Chatellier, et al. 2000)**. Finally, we are not aware of reports that have measured Fas-Ligand and IL-22 on serum or plasma linking their levels to type of infection, and therefore this study is the first, to our knowledge, to report their potential relevancy. Taken together, these findings underscore the need for in-depth studies at the species level. Such analyses are beyond the scope of this study, but are strongly warranted, particularly in type II NSTIs, which are predominantly also caused by other β-hemolytic streptococci, such as *S. dysgalactiae* **(Bruun, Rath, et al. 2021)**.

Our results demonstrate a distinctive inflammatory profile in the different clinical phenotypes, likely resulting from pathogen-specific underlying mechanisms. This concept was further explored through differential connectivity analyses delineating the interconnections, and magnitudes thereof, for each plasma analyte. The results highlighted distinct networks and hubs dependent on the type of NSTI and septic shock. This analysis shifts the focus toward the relationships between analytes rather than on their levels, making it a useful tool in systems biology for investigating and understanding complex biological data **(Saccenti and Svensson 2020; Rosato, Tenori, Cascante, De Atauri Carulla, et al. 2018)**. The potential of personalised medicine in NSTIs has been emphasised in recent reports **(Peetermans et al. 2020)**, and it is tempting to speculate that the hubs identified represent potential targets for interventions, as the associated network is more likely to be affected. It will be of interest in future studies to explore the role of these key hubs in pathophysiology and as therapeutic targets in NSTI.

In conclusion, in this study, we identified discriminatory biomarkers for NSTI and its clinical phenotypes: (a) soft-tissue necrosis (Thrombomodulin); (b) type I versus type II NSTIs (MMP-9, CXCL-10/IP-10, IL-10); and (c) septic shock versus no shock (IL-6, G-CSF, and S100A8). These biomarkers are promising candidates for improved diagnosis and prognosis, which is highly anticipated in clinical practice to decrease the rate of misdiagnosed cases and improve therapeutic strategies in NSTIs.

6.5 Acknowledgements

We thank the patients and relatives for their consent to participate and the clinical staff at the study sites for their invaluable contribution. Thanks are due also to Kristin Rye at the University of Bergen, Norway, for expert technical assistance performing the Luminex analysis. The microbiological laboratories of the participating hospitals are gratefully acknowledged for performing routine culture and identification of microbial etiologies. This work was supported by the Center for Innovative Medicine (CIMED) and Region Stockholm (no. 20180058); the Swedish Research Council (2018-02475); the European Union Seventh Framework Programme (FP7/2007-2013) under



Publication

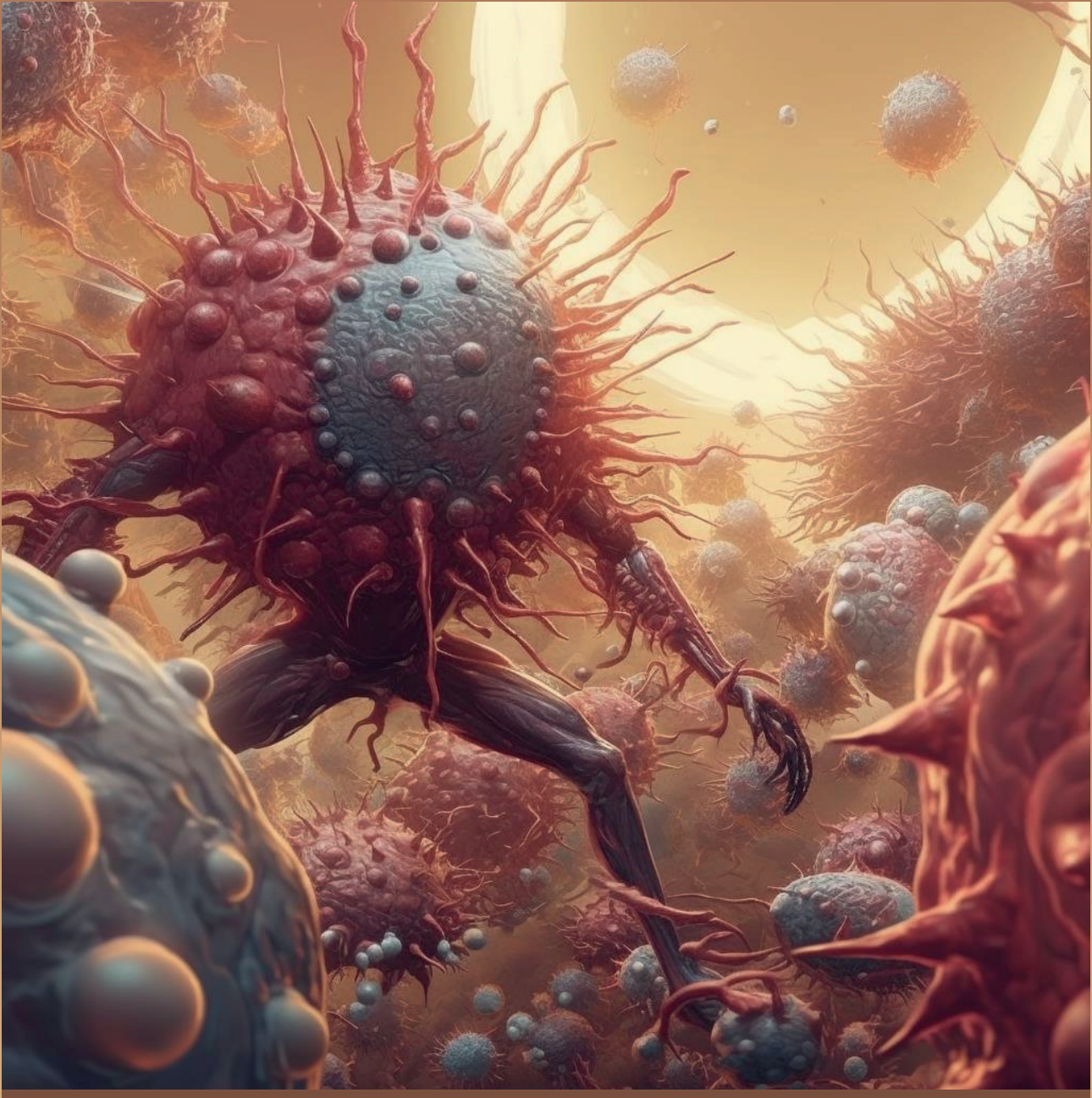
the grant agreement 305340 (INFECT project); the Swedish Governmental Agency for Innovation Systems (VINNOVA), Innovation Fund Denmark, and the Research Council of Norway under the frame of NordForsk (project no. 90456, PerAID); the Swedish Research Council, Innovation Fund Denmark, the Research Council of Norway, the Netherlands Organisation for Health Research and Development (ZonMW), and DLR Federal Ministry of Education and Research, under the frame of ERA PerMed (project 2018-151, PerMIT); and the Swedish Children's Cancer Foundation (TJ2018-0128).

6.6 Author contributions

LMPM, ER, TB, SS, MS, and ANT conceived the project. LMPM, ER, TB, MBM, KS, CU, MBH, PA, MN, OH, SS, MS, and ANT established the protocols for handling clinical material and performing measurements. ER, TB, MBM, MBH, OH, KS, CU, PA, MN, and SS provided resources. LMPM, ER, and ML performed experiments. LMPM was responsible for data curation. LMPM, ER, SJ, TB, and ES carried out formal analysis of the results. LMPM and SJ prepared visualizations. LMPM and ANT wrote the original draft. LMPM, ER, SJ, TB, MBM, KS, CU, MBH, PA, MN, OH, ML, VAPMDS, ES, SS, MS, and ANT reviewed and edited the manuscript. TB, VAPMDS, ES, SS, MS, and ANT supervised the project. OH, VAPMDS, SS, KS, MS, and ANT administered the project, including funding acquisition.

Chapter 7

Chapter



Eivind Rath[2], Laura M. Palma Medina[4*], Sanjeevan Jahagirdar[1*], Knut Anders Mosevoll[2,3], Jan K. Damas[16,17], Martin B. Madsen[5], Mattias Svensson[4], Ole Hyldegaard[5,6], Vitor A. P. Martins dos Santos[7,8], INFECT Study group[#], Edoardo Saccenti[1], Anna Norrby-Teglund[4], Steinar Skrede[2,3], Trond Bruun[2,3]

***Contributed equally #INFECT study group(Oddvar Oppegaard, Torbjørn Nedrebø, Morten Hedetoft, Michael Nekludov)**

Turn to page 377 for author affiliations

This chapter is adapted from:

Rath, E., Palma Medina, L. M.*, Jahagirdar, S.*, Mosevoll, K. A., Damås, J. K., Madsen, M. B., Svensson, M., Hyldegaard, O., Martins dos Santos, V. A. P., Saccenti, E., Norrby-Teglund, A., Skrede, S., & Bruun, T. (2023). Systemic immune activation profiles in streptococcal necrotizing soft tissue infections: A prospective multicenter study. *Clinical Immunology*, *249*, 109276. https://doi.org/10.1016/j.clim.2023.109276

***Contributed equally**

The Immune System Responds: Systemic Immune Activation Profiles

Abstract

Background: Streptococcal necrotizing soft tissue infections (NSTIs) require prompt surgical and medical intervention. Early streptococcal NSTI is often difficult to discern from cellulitis, which primarily requires antibiotics. Increased insight into inflammatory responses along the spectrum of streptococcal disease may lead to identification of new diagnostic targets.
Methods: Hundred-and-two patients with β-hemolytic streptococcal NSTI were derived from the INFECT-study, a prospective Scandinavian multicentre study of NSTIs. Plasma levels of 37 mediators together with leucocytes and CRP were compared to those of 23 cases of streptococcal cellulitis. Associations to infection category, streptococcal species, septic shock and outcomes were investigated.
Results: Differences in mediator levels between NSTIs and cellulitis cases were revealed, in particular for IL-1β, TNFα and CXCL-8/IL-8, all showing AUC values above 0.90. Across streptococcal NSTI etiologies, eight biomarkers separated cases with septic shock from those without, and four mediators predicted a severe outcome. Hierarchical cluster analysis uncovered grouping of biomarker levels corresponding to the type of infection, streptococcal species and outcomes.
Conclusion: Several inflammatory mediators and wider profiles were identified as potential biomarkers of NSTI. Also, NSTI severity was clearly associated to inflammatory profile. Biomarkers for diagnosis and treatment guidance in streptococcal NSTIs are needed and should be further explored.

7.1 Background

Necrotizing soft tissue infections (NSTIs) have high rates of morbidity and mortality **(Hua et al. 2022)**. *S. pyogenes* (group A streptococcus; GAS) is the major cause of monomicrobial (type 2) NSTIs **(M. B. Madsen, Skrede, et al. 2019)**, while *S. dysgalactiae* (SD) is an emerging etiologic cause **(Bruun, Kittang, et al. 2013; Oppegaard et al. 2015)**. Streptococcal NSTIs are frequently associated with bacteremia, septic shock, organ failure and death **(Bruun, Rath, et al. 2021; D. L. Stevens and Bryant 2017)**. Although GAS and SD are the major microbial etiologies also of cellulitis, systemic and local severity signs are less prominent, surgery is rarely needed and complication rates are low **(Gunderson et al. 2018; Bruun, Oppegaard, et al. 2016)**. Of concern, early NSTIs may be mistaken for cellulitis **(Hua et al. 2022; D. L. Stevens and Bryant 2017; Goh et al. 2014; Al Alayed et al. 2015)**. This poses a major threat to many NSTI patients, as it may delay surgery and antibiotic therapy, the most important measures to reduce organ failure, sequelae and death **(Miranda et al. 2018; Hadeed et al. 2016; Nawijn et al. 2020)**. The diagnosis of NSTI is still based on surgical exploration and confirmation **(Goh et al. 2014)**.

Cytokine profiles in sepsis are well characterised **(Van Engelen et al. 2018; Pierrakos et al. 2020)**, but studies have failed to identify plasma mediators that reliably diagnose or prognosticate septic shock and mortality **(Van Engelen et al. 2018; Pierrakos et al. 2020)**. Therefore, combinations of mediators have been advocated **(Van Engelen et al. 2018; Pierrakos et al. 2020)**. Cytokines have also been explored as diagnostic and prognostic biomarkers in specific infectious diseases **(Struck et al. 2020; Wright et al. 2021)**. Apart from streptococcal toxic shock syndrome (STSS) **(Low 2013; Norrby-Teglund, Thulin, et al. 2001)**, inflammatory profiles of streptococcal NSTIs and cellulitis are insufficiently described **(Hansen, Rasmussen, Svensson, et al. 2017; Saccenti and Svensson 2020; Kristensen et al. 2020)**.

Improved biomarker-based tools are needed to support the clinician in establishing the diagnosis of NSTIs at an early stage of the disease. For streptococcal skin and soft tissue infections (SSTIs) in particular, there is a wide and overlapping spectrum of disease that requires investigation of differential inflammatory patterns. This may advance the understanding of the pathogenesis, lead to development of new diagnostic and prognostic tools and possibly identify potential targets of treatment. In a large Scandinavian multicenter patient cohort of NSTIs, we recently demonstrated that inflammatory profiles differ between polymicrobial and monomicrobial infections and by the presence of shock **(Medina et al. 2021)**. In the present study, we explore the systemic immune activation pattern of the streptococcal NSTI cases, comparing and contrasting them to streptococcal cellulitis.

7.2 Methods

7.2.1 Setting & patients

This work was conducted as part of the INFECT-study, a Scandinavian multicenter study, with a prospectively included cohort of 409 NSTI cases in total (ClinicalTri-

als.gov (NCT01790698)) **(M. B. Madsen, Skrede, et al. 2019)**. Cases caused by GAS (n = 126) and SD (n = 27) were selected, and of the 153 streptococcal cases identified **(M. B. Madsen, Skrede, et al. 2019; Bruun, Rath, et al. 2021)**, 102 were strictly monomicrobial and included for further analysis in this study.

Two control groups with a total of 43 cases were included. The cellulitis control group consisted of four operated (surgery disproved NSTI), and 19 non-operated cellulitis cases with confirmed or probable GAS or SD aetiology defined by positive culture or serology as previously described **(Bruun, Oppegaard, et al. 2016)**. The healthy control group consisted of patients admitted for elective orthopaedic surgery (n = 20), with blood samples taken prior to surgery (Copenhagen University Hospital, Denmark, Regional ethics committee permit: H-2-2014-071). Ethical approval and patients' consent are as described previously **(M. B. Madsen, Skrede, et al. 2019; Hansen, Rasmussen, Svensson, et al. 2017)**.

7.2.2 Clinical characteristics

Clinical and laboratory data were acquired according to the INFECT-study protocol **(M. B. Madsen, Skrede, et al. 2019)**. Outcomes registered in the NSTI cohort included septic shock at admission, along with death and/or amputation. Septic shock was defined as use of vasopressor or inotrope and lactate ≥ 2 mmol/L **(M. Madsen et al. 2018)**. Severe outcome was defined as amputation and/or death within 90 days from admission. Analysing this outcome, we restricted inclusion to cases with an extremity as a primary site of infection and excluded cases amputated prior to arrival at study hospital.

7.2.3 Plasma biomarker analysis

Blood was collected into 10mL EDTA vacuum tubes and kept on ice until processed (15-90 minutes max). Plasma was collected after centrifugation (2500G, for 10minutes at 20°C), aliquoted into cryotubes and stored at −80°C until further analysis. A 32 and a 5-plex customized multiplex assay (Magnetic Luminex Assay, R&D Systems, Inc; Abingdon, UK) were used for analyzing the following mediators, interleukins: IL-1α, IL-1β, IL-1RA, IL-2, IL-4, IL-6, IL-10, IL-12p70, IL-13, IL-17A, IL-18, IL-22, IL-36β/IL-1F8; chemokines: CCL-2/MCP-1, CCL-4/MIP-1β, CCL-5/RANTES, CXCL-8/IL-8, CXCL-10/IP-10; soluble adhesion molecules: E-Selectin, ICAM-1, VCAM-1; matrix metalloproteases: MMP-1, MMP-8, MMP-9; growth factors: G-CSF; and molecules designated "others": C5/C5a, Collagen-IVα1, I-α-1/COL1A1, Fas-Ligand, Galectin-3, MPO, Pentraxin-3, Resistin, S100A8, S100A9, Thrombomodulin, and TNFα. In addition, IL-23 and IL-33 were analyzed using ELISA technique (R&D Systems; Abingdon, UK), whereas C-reactive protein (CRP) and leucocytes were analyzed according to laboratory diagnostic routines. Analyte concentration measurements were performed according to manufacturer's instructions. An overview of the analyzed mediators and mode of analysis is detailed in Table 7.1. For concentrations outside their respective detection ranges (out of range, OOR), an imputation strategy was applied, as described previously **(Medina et al. 2021)**. Due to high rates of OOR measurements, the analytes IL-1RA and IL-33 were omitted for further analysis. For further details see section *Imputation strategy for left and right-censored mediator data*. In total,

37 systemic plasma mediators together with leukocytes and CRP were included in the analyses.

7.2.4 Imputation strategy for left and right-censored mediator data

We initially included a total of 39 biomarkers. However, IL-33 had > 50% Out of Range (OOR) below values and IL-1RA had 40% OOR values and both were excluded for further analysis. CXCL-10/IP-10 showed 29% OOR above, but was kept for the final analysis after the imputation method was applied. Imputation of Left ($OOR <$) and Right ($OOR >$) censored data was performed using the method proposed by Wei et al. **(R. Wei et al. 2018)**. With this imputation approach constraints are used so that the imputed values are larger than the maximum observed (for $OOR >$ censored) or smaller than the minimum observed (for $OOR <$ censored). Two mediators had only one $OOR >$ measurement, i.e. E-Selectin and IL-10. These have been substituted to $max + (0.2 \times max)$ (min observed, including extrapolated values). After this initial processing, two cytokines were doubly censored: S100A9 and IL-23. A two-step approach was taken to deal with these cases. Data was made left-censored by setting the $OOR >$ censored to the maximum observed value and then the left-censored were imputed with the proposed strategy. After imputation the $OOR >$ previously set to max were set back to censored and imputed again.

In studies on biomarker concentration, different approaches with respect to handling values OOR (i.e. censored) have been used. Low values (left censored data, or $> OOR$) have often been handled in one of three manners: (i) set to half of the lowest value measured, (ii) lowest value divided on the $\sqrt{2}$, or (iii) set to the lowest measured level (or even zero) **(R. Wei et al. 2018)**. Samples with values OOR above should ideally be re-analysed after diluting the sample. However, we did not acquire sufficient volume of plasma to perform a second analysis. Another approach used in such a setting is to set the OOR values above the highest measured value. Our imputation strategy offers a reasonable approach with even distribution of the OOR values at the correct end of the scale, hence reducing statistical bias compared to 'set-value approaches' formerly frequently used.

The resampling method we applied creates an artificial larger population based on the included cases that results in more stringent p-values. In all the analyses performed, more significant findings and lower p-values were seen when the analyses were done without resampling. In all, by these methodological approaches it is ensured that the findings produced are markedly less exposed for type 1 errors.

7.2.5 Experimental model: In vitro stimulation

The level of selected mediators was measured in an in vitro stimulation experiment using human peripheral blood mononuclear cells (PBMC) isolated from healthy blood donors exposed to supernatants and heat-inactivated GAS and SD isolates from NSTI cases, and in human umbilical vein endothelial cells (HUVEC) exposed to supernatants from the bacterial-stimulated PBMCs.

Table 7.1: Overview: Biomarkers analyzed and reference value intervals a IL-1RA and IL-33 omitted because of high numbers of out-of-range values. b According to the manufacturer (R&D Systems). c Analysed by ELISA technique.

Biomarkers[a]	Reference interval[b] (pg/ml)
Interleukins	
IL-1α	5.3-1290
IL-1β	19.5-4740
IL-2	31.1-7560
IL-4	15.7-3820
IL-6	4.7-1150
IL-10	4.8-1160
IL-12p70	133.6-32460
IL-13	367.8-89370
IL-17A	12.5-3030
IL-18	10.1-2460
IL-22	14.3-3470
IL-23 [c]	125-8000
IL-36β/IL-1F8	3.95-960
Chemokines	
CCL-2/MCP-1	33-8020
CCL-4/MIP-1β	150-36490
CCL-5/RANTES	22.6-5490
CXCL-8/IL-8	5.2-1260
CXCL-10/IP-10	2.1-510
Adhesion molecules	
E-Selectin	336.1-81670
ICAM-1	6975.6-1695060
VCAM-1	8233-2000610
Matrix metalloproteases	
MMP-1	49.8-12090
MMP-8	245.1-59560
MMP-9	134.2-32600
Growth factors	
G-CSF	22.7-5520
Other	
C5/C5a	843.6-205000
Pentraxin-3	206.8-50250
TNFα	9.7-2360
S100A8	74.9-18190
S100A9	26.9-6530
MPO	127.9-31070
Fas-Ligand	9-2190
Thrombomodulin	84.4-20500
Galectin-3	16.8-4070
Collagen-IVα1	93-22600
I-α-1/COL1A1	13-3170
Resistin	72.6-17650

Bacterial culture

Two bacterial strains isolated from patients included in the INFECT study were used for in vitro stimulation of peripheral blood mononuclear cells (PBMCs). The bacterial isolates from patients denoted 2006 and 6017 were selected as representatives of GAS and SD, respectively. Liquid bacterial cultures were made in Todd Hewitt media (THW) with 1.5% (wt/vol) of yeast extract from a single bacterial colony in blood agar plates. Serial dilutions were made from the primary culture and incubated at $37°C$. After 14-16 h incubation, 3 ml of cultures in stationary and exponential phase were collected. The bacterial pellet was obtained by centrifugation at 4000 rpm for 5 min. The supernatant was removed and filtered with a 0.2 μm filter, while the bacterial pellet was washed twice with Phosphate Buffered Saline (PBS) and thereafter killed by incubation at $75°C$ for 30 min.

Cell culture

The PBMCs were isolated by Ficoll-Hypaque density gradient centrifugation (Lymphoprep, Axis-Shield). The cells were cultured in RPMI-1640 supplemented with 10% FBS (Sigma-Aldrich), 2 mM L-glutamine (Thermo Fisher Scientific), and 25 mM HEPES (Thermo Fisher Scientific). The cells were kept overnight at $4°C$ before being seeded in 48-well plates for further use. Human umbilical vein endothelial cells (HUVEC) were cultured in completed endothelial cell growth medium 2 (Sigma-Aldrich) at $37°C$ in a 5% CO_2 incubator. A cryostock of HUVEC with 5×10^6 cells was dissolved in 30 ml media and cultured until confluence in a T-175 flask. Cells were detached by trypsination for 5 min at $37°C$ degrees. For seeding, the cells were resuspended in media at a concentration of 2×10^5 cells/ml and 200 μl were seeded per well on a 96-well plate and incubated for 2 days prior to stimulation.

PBMC and HUVEC stimulation

The stimulation assay consisted of two sequential experiments. First, the PBMCs were stimulated for 24 h with supernatant or heat-inactivated cells of GAS and SD strains, isolated from NSTI cases. Then the resulting supernatants from this incubation were used to stimulate HUVEC for 24 h. PBMCs isolated the day before were seeded at a concentration of 2×10^6 per well in a 48-well plate and rested for 2 h at $37°C$ in a 5% CO2 incubator. The stimulation of PBMC was carried out with a mix of stationary and exponential bacterial supernatant diluted 1:20 or heat-killed bacteria at an equivalent multiplicity of infection (MOI) of 3. The cells were then incubated for 24 h at $37°C$ in a 5% CO2 incubator. After the incubation period, the cell culture supernatant was collected for analysis, but also to stimulate HUVEC cells that have reached confluency. Stimulation of HUVEC cells was performed with the PMBC media diluted 1:4 in HUVEC media. Following this, cells were then incubated for another 24 h at $37°C$ in a 5% CO2 incubator, and media was collected for measurements. All collected media from the PBMC and HUVEC stimulation was initially frozen at -$20°C$ overnight and then passed to -$80°C$ for long-term storage until measurement.

Measurement of analytes in cell culture media after stimulation was performed using customized Luminex assays, following the instructions of the manufacturer (R&D). Different panels of analytes were used for each cell media supernatant. PBMC

media was tested for IFN-γ, TNFα and IL-1β, while CXCL-10/IP-10, S100A9, ICAM-1 and E-Selectin were measured in the media from the HUVEC stimulations.

The data from the in vitro stimulation tests were analyzed in GraphPad Prism version 8.0.0 for Windows, using Wilcoxon matched-pairs signed rank test.

7.2.6 Statistical analysis

Analyses were done using R programming language (R Core Team, 2019) and IBM SPSS Statistics for Windows, version 24.0 (IBM Corp., Armonk, NY). For categorical variables Fisher's exact test or $\chi 2$ test were used as appropriate. Due to the non-normality of the data, Mann-Whitney U test was applied for continuous variables. Due to the unequal group size of the clinical categories compared, a resampling procedure (iteration) of the data results was performed. The resampling included 10^4 iterations based on 90% of the smallest group size used for each comparison. An average p-value of all comparisons was used to evaluate the final results.

All tests were two-sided and differences were considered significant at a p-value < 0.05. Due to high numbers of mediators evaluated, a Benjamini-Hochberg adjustment for multiple testing was applied **(Benjamini et al. 1995)**.

The discriminant ability of the most significant biomarker differences was evaluated through receiver operation characteristic (ROC) curves and the corresponding area under the ROC curve (AUC) was calculated. For correlations analysis Spearman's correlation (rho, ρ) was calculated.

Additionally, we performed unsupervised hierarchical cluster analysis, where the cases and biomarkers are categorized based on relatedness and combined with a heat map to visualize correlations. The mediators were log_{10} transformed and converted to z-score **(Cheadle et al. 2003)**, before analysis using J-Express **(Stavrum et al. 2008)**. Euclidean distance and complete linkage were used.

7.2.7 RandomForests

Random Forest (RF) classification was used to build classification models to discriminate among cytokine profiles of patients having: 1) streptococcal cellulitis and NSTIs; 2) NSTIs of GAS and SD aetiology; and to assess 3) severe manifestations and the association of amputation and death (designated "severe outcome"). Key clinical variables (i.e. gender, age, septic shock, HBOT, IVIG) were also included (specified in the footnote of the different tables).

The RF classification model is based on growing an ensemble of decision trees and counting each tree's classification vote to determine the most popular class **(Breiman 2001)**. Each decision tree is grown from random vectors generated by bootstrapped datasets, that consists of randomly selected samples from the original dataset: This process of bootstrapping and aggregating the votes of the decision trees uses $\frac{1}{3}$ of the samples to build the actual model and leaves out $\frac{1}{3}$ of the samples for model validation.

The building of these decision trees is governed by split selection, using a random subset of variables (systemic mediators, in this case) and considering the Gini impurity. The quality of the prediction/classification RF model (classification error)

is assessed on the unused samples and is known as Out-of-Bag (OOB) Error. This embedded cross-validation procedure ensures unbiased evaluation of the classification model since the model quality is evaluated on samples which have not been used to train the model. All RF classification models were constructed as described in the statistical analysis section of the article. RF models were built using the R package rfPermute **(Archer et al. 2016)**. Statistical significance was assessed by means of permutation test as implemented in the R package rfPermute using 100 permutation of the original data set. All calculations were performed in R **(R Core Team 2013)**.

We use the RF classification models in this study to evaluate the cumulative effects of all cytokines in a multivariate approach. The multivariate approach could yield different results compared to univariate statistics that consider the effect of one cytokine at a time as the multivariate approach often compensates for complex interactions between various cytokines. As this approach is relatively robust against differences in sample size and mitigates issues that may arise from inconsistencies in the data, we believe that the benefits of using such an approach in tandem with the other approaches is high.

7.2.8 Network and differential connectivity analysis

Network of associations between mediators were built, using the Probabilistic Context Likelihood of Relatedness on Correlation (PCLRC), as described elsewhere **(Jahagirdar, Suarez-Diez, et al. 2019)**. PCLRC gives a measure of association and the probability of likelihood in occurrence of the relationship between the mediators. Associations with a weighted probability > 0.95 were kept in the analysis. The weighted probability is interpreted as a confidence level based on which we can accept or reject the association between the pair of mediators. Different networks were built separately for GAS and SD cases, with and without septic shock

Differential connectivity analysis was applied to compare the mediators in the association networks. Connectivity can be interpreted as how much one mediator is significant or affects the entire system (of all mediators) as it takes into account both the number and strength of connections. Differential connectivity can be interpreted as the difference in the significance or effect of one mediator in the two situations. The higher the difference, the higher change in the role that mediator plays in the two comparing situations. The procedure was performed as described elsewhere **(Medina et al. 2021; Jahagirdar and Saccenti 2020b)**.

7.2.9 Data availability

The entire raw data with and without imputed values are available at
DOI: 10.5061/dryad.f1vhhmgw4.

7.3 Results

7.3.1 Clinical characteristics

Among the 102 monomicrobial streptococcal NSTI cases, 88 were caused by GAS and 14 by SD. Demographics and characteristics of these, and the 23 streptococcal

cellulitis controls, are summarized in Table 7.2. Comorbid conditions were prevalent among both NSTI and cellulitis patients, and most infections were located in the extremities. Bacteremia, septic shock, treatment in ICU, and death were more frequent among NSTI cases. Details on risk factors, pre- and preoperative findings, treatment including time from admission to primary surgery and total number of operations per patient in the NSTI cohort have been described previously **(Bruun, Rath, et al. 2021)**.

7.3.2 Mediator levels in streptococcal necrotizing soft tissue infections and cellulitis

In total, 29 of 37 plasma mediators, in addition to CRP, were significantly elevated in the NSTI compared to the cellulitis cases, of which 28 showed an AUC above 0.80 (Table 7.3). In the multivariate RF model adjusting for age, gender and septic shock, a set of nine mediators differentiating NSTI and cellulitis were identified, based on discriminatory power according to Gini index values. Predictors identified in the model were IL-1β, TNFα, CXCL-8/IL-8, MMP-8, IL-6, Pentraxin-3, IL-22, CCL-4/MIP-1β and S100A8, that all displayed AUC values > 0.86. Additionally, a comparison between NSTI and cellulitis cases without septic shock was performed, in which all the same mediators were predictors in the RF model, with the exception of IL-6 (Table 7.3). In this latter comparison, CRP also showed a significant result, in both the univariate and multivariate analyses. Several of the plasma mediators with significant findings in the RF model even displayed higher AUC in this latter comparison.

Comparing cellulitis cases and healthy controls, 18 mediators were significantly higher in cellulitis, whereas three mediators were more elevated among the controls (Table 7.5).

7.3.3 Mediator levels associated to severity and outcome in necrotizing soft tissue infections

In streptococcal NSTIs with septic shock at admission, 25 mediators were significantly elevated compared to cases without shock. Opposite, the concentration of MMP-9 was higher in the group without shock. The RF model, with adjustment for age and gender, identified eight independent relevant predictors of shock (high Gini index), with an associated AUC greater than 0.80; IL-4, IL-6, IL-36β/IL-1F8, CCL-2/MCP-1, CXCL-8/IL-8, G-CSF, Pentraxin-3 and S100A8 (Table 7.6).

A significant positive correlation was seen between Sequential Organ Failure Assessment (SOFA) score at the day of admission and 30 out of 37 plasma mediators, but not CRP or leucocytes. Two mediators had a significant negative correlation (Table 7.7).

No associations of severe outcome and mediators were detected by univariate analysis. In the RF model, four mediators (IL-6, IL-10, G-CSF and Collagen-IVα1) were associated with severe outcomes, all with an AUC value above 0.70 (Table 7.8).

Table 7.2: Demographics, clinical characteristics and outcomes in patients with streptococcal NSTIs and cellulitis. Abbreviations: NSTIs: Necrotizing soft tissue infections, GAS: *S. pyogenes* (Group A streptococcus), SD: *S. dysgalactiae*, BMI: Body mass index, BHS: β-hemolytic streptococci, IVIG: Intravenous immunoglobulin, ICU: Intensive care unit, HDU: High dependency unit, SOFA: Sequential Organ Failure Assessment score, SAPS II: Simplified Acute Physiology Score. The data are given as median values with interquartile range (IQR) and numbers, percentages in parentheses. a Represents comparison of all NSTI cases vs cellulitis cases (either χ2test or Mann-Whitney U test, as appropriate). b Missing two in the GAS cohort. c Active malignancy, chronic obstructive pulmonary disease or asthma, current or previous cardiovascular disease, diabetes mellitus, chronic kidney failure, chronic liver disease, rheumatoid disease, immunodeficiency/immunosuppression. d Intake as defined by **(M. Madsen et al. 2018)**. e Including also those amputated before admission to the study hospital. Amputation is defined as the surgical removal of all or part of a limb or extremity. f Missing data for two cases. g Missing data for one case. h Data is only available for four of the cellulitis cases (i.e. the four cases with suspected NSTI, in whom surgery later showed cellulitis). i Statistical comparison was not performed due to a low number of cases with available data in the cellulitis cohort.

	GAS NSTIs (n = 88)	**SD NSTIs** (n = 14)	**All NSTIs** (n = 102)	**Cellulitis** (n = 23)	**p-value** [a]
Demographics					
Age (years)	59.5 (48-69)	67.5 (60-73)	60.5 (48-70)	43.0 (38-62)	0.009
Sex, male gender	44 (50)	8 (57)	52 (51)	16 (70)	0.106
BMI[b]	25.9 (23.4-30.2)	25.6 (23.1-28.1)	25.9 (23.3-29.4)	27.7 (23.43-33.8)	0.233
Underlying condition					
Significant comorbidity[c]	52 (59)	12 (86)	64 (63)	12 (52)	0.348
Active smoker	15/77 (19.5)	3/12 (25)	18/89 (20)	3/19 (16)	1
High alcohol consumption[d]	6/63 (9.5)	4/11 (36)	10/74 (13.5)	0/19 (0)	0.205
Outcome variables					
Blood culture positive BHS	46 (52)	8 (57)	54 (52.9)	3 (13)	0.001
Septic shock (at baseline)	55 (62.5)	9 (64.3)	64 (63)	3 (13)	<0.001
IVIG treatment	64 (73)	8 (57)	72 (70.6)	0 (0)	<0.001
Amputation[e]	11 (12.5)	3 (21.4)	14 (13.7)	0 (0)	0.07
Mortality (day 90)	8 (9)	5 (36)	13 (12.7)	0 (0)	0.124
Mortality and/or amputation	16 (18.2)	7 (50)	23 (22.5)	0 (0)	0.007
SOFA score, day 1	9 (6-12)[f]	11 (8-13)[g]	9 (6-12)	8 (5-13)[h]	NA[i]
SAPS II, day 1	40 (33-55)[f]	56 (47-63.5)[g]	42 (34-58)	32 (23.5-48.5)[h]	NA[i]
Hospitalization at					
ICU/HDU	88 (100)	14 (100)	102 (100)	4 (17.4)	<0.001
Primary site of infection					
Head/neck	11 (12.5)	1 (7.1)	12 (11.8)	9 (39.1)	NA
Upper extremities	38 (43.2)	1 (7.1)	39 (38.2)	5 (21.8)	NA
Lower extremities	34 (38.6)	12 (85.8)	46 (45.1)	9 (39.1)	NA
Abdomen/anogenital	5 (5.7)	0 (0)	5 (4.9)	0 (0)	NA

Table 7.3: Biomarker levels in streptococcal NSTIs vs cellulitis. Abbreviation: NSTIs: Necrotizing soft tissue infections. Unit of measurement: pg/mL. Data are presented as median values with interquartile ranges (IQR). a Calculated after resampling as described in the methods section. b Mann-Whitney U test p-values after Benjamini-Hochberg adjustment: * ≤0.05; ** ≤0.01; *** ≤0.005. NS: non-significant. c AUC: Area under receiver operating curve. AUC values are presented exclusively for biomarkers with statistically significant differences obtained after Benjamini- Hochberg adjustment. d RF: Random forest. No. of trees: 100.000. Split: 6. Repetitions: 100. Accuracy: 92.7%. Age, gender and septic shock are clinical parameters included in the RF modelling. e Unit of measurement: $\times 10^9$/L. Missing six in the GAS cohort. f Unit of measurement: mg/L. Missing five in the GAS cohort.

	NSTIs	Cellulitis	Mann-Whitney U[a]	AUC[c]	RF[d]	
	n = 102	n = 23	p-value[b]		Gini index	p-value
Interleukins						
IL-1α	54 (46-64)	30 (21-40)	***	0.893	0.37	1
IL-1β	17 (12-30)	0.2 (0.2-1.9)	***	0.91	3.37	<0.01
IL-2	1011 (561-1367)	402 (236-626)	NS	NA	0.25	1
IL-4	227 (184-263)	120 (98-151)	***	0.892	0.34	1
IL-6	775 (185-7523)	20 (15-78)	***	0.894	1.85	<0.01
IL-10	70 (40-116)	13 (6-36)	*	0.813	0.29	1
IL-12p70	197 (127-263)	22 (0.6-88)	***	0.835	0.46	1
IL-13	1464 (1154-1688)	738 (603-950)	***	0.905	0.41	1
IL-17A	30 (16-70)	7 (3- 11)	***	0.851	36	1
IL-18	566 (372-1024)	341 (270-490)	*	0.732	0.35	1
IL-22	120 (100-139)	62 (42-77)	***	0.906	1.49	0.03
IL-23	1857 (307-7146)	622 (37-10,026)	***	0.582	1.75	0.06
IL-36β/IL-1F8	18 (15-20)	8 (8-12)	***	0.892	0.91	0.27
Chemokines						
CCL-2/MCP-1	868 (380-2050)	153 (119-271)	***	0.882	1.01	0.22
CCL-4/MIP-1β	880 (763-989)	525 (488-626)	***	0.919	1.45	0.03
CCL-5/RANTES	6575 (2332-12,480)	5104 (1630-9727)	NS	NA	0.57	0.97
CXCL-8/IL-8	43 (18-226)	6 (4-13)	***	0.92	2.2	<0.01
CXCL-10/IP-10	5001 (617-383,052)	282 (177-916)	NS	NA	0.36	1
Adhesion molecules						
E-Selectin	145171 (97983-208659)	42453 (33220-67367)	***	0.861	0.44	1
ICAM-1	667966 (544604-876826)	355418 (224411-518623)	**	0.821	1.16	0.21
VCAM-1	43×10^5 (33×10^5-63.1×10^5)	29.4×10^5 (12.7×10^5-34.3×10^5)	**	0.803	0.26	1
Matrix metalloproteases						
MMP-1	1566 (973-3112)	620 (453-877)	***	0.85	0.59	0.96
MMP-8	31,555 (11,820-65,403)	1846 (758-6632)	***	0.93	2.1	<0.01
MMP-9	5527 (3226-16,509)	8145 (6358-13,664)	NS	NA	1	0.48
Growth factors						
G-CSF	2465 (437-24,977)	155 (128-314)	***	0.848	0.65	0.86
Other						
C5/C5a	21,703 (14,176-32,769)	30,665 (13,715-60,214)	NS	NA	0.64	0.95
Pentraxin-3	15,993 (6549-26,957)	880 (478-5373)	***	0.884	1.92	0.03
TNFα	35 (22-58)	10 (7-12)	***	0.938	2.56	<0.01
S100A8	1014 (690-1799)	328 (213-528)	***	0.864	1.26	0.04
S100A9	2628 (1508-4569)	568 (450-1349)	***	0.816	1.18	0.29
MPO	41,121 (31,004-55,724)	20,150 (15,294-26,418)	**	0.838	0.26	1
Fas-Ligand	36 (26-53)	36 (23-47)	NS	NA	0.4	1
Thrombomodulin	12,731 (9028-16,146)	5764 (4919-7030)	***	0.908	1.87	0.06
Galectin-3	3318 (2771-3956)	2524 (2208-3169)	NS	NA	0.53	0.99
Collagen-IVα1	2549 (1632-3745)	940 (610-1247)	***	0.902	78	0.73
I-α-1/COL1A1	9552 (7104-19,088)	6457 (4605-10,049)	NS	NA	0.55	0.98
Resistin	55,062 (40,589-64,903)	19,083 (14,410-39,174)	***	0.818	0.27	1
Leucocytes[e]	14.3 (8.1-20.6)	9.6 (7.1-12.5)	NS	NA	0.52	1
CRP[f]	272 (191-361)	54 (36-151)	***	0.814	1.52	0.21

Table 7.4: Biomarker levels in streptococcal NSTIs vs cellulitis, without septic shock. Abbreviation: NSTIs: Necrotizing soft tissue infections. Unit of measurement: pg/mL. Data are presented as median values with interquartile ranges (IQR). a Calculated after resampling as described in the methods section. b Mann-Whitney U test p-values after Benjamini-Hochberg adjustment: *≤0.05; **≤0.01; ***≤0.005. NS: non-significant. c AUC: Area under receiver operating curve. AUC values are presented exclusively for biomarkers with statistically significant differences obtained after Benjamini-Hochberg adjustment. d RF: Random forest. No. of trees: 100000. Split: 6. Repetitions: 100. Accuracy: 92.5%. Age and gender are clinical parameters included in the RF modelling. e Unit of measurement: $\times 10^9$/L. Missing one in the NSTI cohort. f Unit of measurement: mg/L. Missing one in the NSTI cohort.

	NSTIs	Cellulitis	Mann-Whitney U[a]	AUC[c]	RF[d]	
	n = 38	n = 20	p-value[b]		Gini index	p-value
Interleukins						
IL-1α	46 (39-50)	29 (20-35)	***	0.87	0.25	1
IL-1β	12 (7-17)	0.2 (0.2-0.2)	***	0.95	3.89	<0.01
IL-2	583 (411-1017)	384 (158-574)	NS	-	0.2	1
IL-4	182 (152-213)	116 (89-139)	***	0.896	0.39	1
IL-6	216 (86-558)	19 (12-57)	***	0.908	0.7	0.5
IL-10	44 (18-67)	12 (5-24)	*	0.791	0.14	1
IL-12p70	131 (75-196)	17 (0.6-57)	***	0.875	0.27	1
IL-13	1218 (955-1474)	649 (580-801)	***	0.867	0.27	1
IL-17A	19 (12-30)	5 (2-9)	***	0.836	0.27	1
IL-18	546 (291-856)	343 (276-522)	NS	-	0.13	1
IL-22	101 (87-121)	54 (40-66)	***	0.934	1.3	<0.01
IL-23	1460 (196-4961)	424 (4-5575)	***	0.604	3.63	<0.01
IL-36β/IL-1F8	15 (12-16)	8 (7-9)	***	0.909	0.92	0.1
Chemokines						
CCL-2/MCP-1	387 (210-594)	138 (111-168)	***	0.842	0.54	0.94
CCL-4/MIP-1β	767 (699-835)	515 (459-582)	***	0.943	1.35	0.02
CCL-5/RANTES	7490 (3684-12088)	6420 (1828-10975)	NS	-	0.22	1
CXCL-8/IL-8	19 (12-31)	6 (4-9)	***	0.913	1.27	0.02
CXCL-10/IP-10	981 (371-11054)	239 (152-514)	NS	-	0.36	1
Adhesion molecules						
E-Selectin	121227 (71708-185216)	40411 (29684-46450)	***	0.899	0.75	0.45
ICAM-1	616286 (527608-771348)	335616 (224411-557381)	*	0.776	0.68	0.73
VCAM-1	39×10^5 (21×10^5-53.1×10^5)	18×10^5 (11×10^5-34×10^5)	*	0.758	0.29	1
Matrix metalloproteases						
MMP-1	1243 (856-1875)	589 (402-777)	***	0.854	0.67	0.78
MMP-8	14578 (8403-35893)	1011 (690-3402)	***	0.922	1.77	<0.01
MMP-9	5527 (3226-16509)	8145 (6358-13664)	NS	-	0.73	0.55
Growth factors						
G-CSF	441 (268-1099)	153 (124-203)	**	0.822	0.27	1
Other						
C5/C5a	22401 (15613-34755)	31359 (13715-67055)	NS	-	0.54	0.96
Pentraxin-3	5735 (3163-12991)	642 (449-2839)	***	0.879	1.27	0.04
TNFα	22 (17-30)	9 (6-11)	***	0.949	1.74	<0.01
S100A8	707 (490-865)	284 (200-407)	***	0.872	1.31	0.02
S100A9	2151 (1477-3395)	568 (450-1349)	***	0.782	0.54	0.96
MPO	31046 (24733-40880)	20920 (13967-26418)	*	0.762	0.15	1
Fas-Ligand	30 (20-38)	28 (21-42)	NS	-	0.28	1
Thrombomodulin	11424 (7955-13082)	5526 (4885-6914)	***	0.866	1.22	0.14
Galectin-3	3058 (2606-3503)	2737 (2323-3246)	NS	-	0.21	1
Collagen-IVα1	1846 (1365-2905)	851 (557-1058)	***	0.891	0.74	0.51
I-α-1/COL1A1	8030 (5302-13384)	6934 (4803-23004)	NS	-	0.19	1
Resistin	46847 (31006-60136)	17782 (14048-32258)	***	0.842	0.32	1
Leucocytes[e]	16.3 (11.9-20.8)	9.6 (6.9-12.0)	**	0.821	0.47	0.93
CRP[f]	270 (200-337)	54 (36-116)	***	0.834	1.74	0.02

Table 7.5: Biomarker levels in cellulitis and healthy controls at admission. Unit of measurement: pg/mL. Data are resented as median values with interquartile ranges (IQR). Leucocytes and C-reactive protein were not included, as samples were not collected from healthy controls. a In order to illustrate the differences between moderate cellulitis and healthy controls, the streptococcal cellulitis cohort in this comparison does not include the four cases with suspected NSTI, in whom surgery later showed cellulitis. b Mann-Whitney U test p-values: *≤0.05, **≤0.01, ***≤0.005 after Benjamini-Hochberg adjustment. NS: non-significant c AUC: Area under receiver operating curve. AUC values are presented exclusively for biomarkers with statistically significant differences obtained after Benjamini-Hochberg adjustment. d Healthy controls > cellulitis.

	Cellulitis[a] n = 19	**Healthy controls** n = 20	**Mann-Whitney U p-value[b]**	**AUC [c]**
Interleukins				
IL-1α	28 (20-33)	19 (14-28)	NS	-
IL-1β	0.2 (0.2-0.2)	0.2 (0.2-0.2)	NS	-
IL-2	385 (158-574)	54 (6-244)	***	0.779
IL-4	112 (89-129)	98 (79-121)	NS	-
IL-6	18 (12-48)	3 (2-8)	***	0.879
IL-10	13 (6-24)	0.2 (0.2-4)	***	0.908
IL-12p70	12 (1-45)	4 (1-52)	NS	-
IL-13	627 (580-780)	439 (338-543)	***	0.789
IL-17A	5 (3-9)	0.1 (0.1-0.1)	***	0.958
IL-18	345 (287-522)	213 (105-257)	***	0.897
IL-22	52 (40-66)	18 (8-29)	***	0.9
IL-23	622 (37-5575)	1 (0.1-322)	**	0.753
IL-36β/IL-1F8	8 (7-9)	7 (5-8)	*	0.716
Chemokines				
CCL-2/MCP-1	137 (111-166)	173 (146-210)	* [d]	721
CCL-4/MIP-1β	509 (459-552)	443 (399-501)	NS	-
CCL-5/RANTES	6250 (1828-9727)	16531 (10093-25398)	*** [d]	0.842
CXCL-8/IL-8	6 (4-8)	6 (3-8)	NS	-
CXCL-10/IP-10	252 (153-514)	76 (54-115)	***	0.874
Adhesion molecules				
E-Selectin	40888 (29684-46450)	27617 (20582-35394)	***	0.768
ICAM-1	355418 (248809-557381)	371233 (283851-758185)	NS	-
VCAM-1	19.5×10^5(12.6×10^5-33.6×10^5)	9.4×10^5 (7.4×10^5-11.3×10^5)	***	0.804
Matrix metalloproteases				
MMP-1	604 (402-777)	536 (376-958)	NS	-
MMP-8	995 (690-2531)	602 (417-817)	*	0.724
MMP-9	8728 (6358-13664)	21993 (14249-38229)	*** [d]	0.797
Growth factors				
G-CSF	151 (124-201)	117 (106-146)	NS	-
Other				
C5/C5a	32053 (16719-67055)	12196 (9042-15867)	***	0.784
Pentraxin-3	637 (450-1679)	383 (58-499)	***	0.816
TNFα	9 (6-11)	12 (7-14)	NS	-
S100A8	263 (200-354)	189 (146-269)	NS	-
S100A9	568 (488-1349)	311 (254-489)	***	0.803
MPO	19844 (13967-25185)	15157 (12753-22610)	NS	-
Fas-Ligand	32 (23-42)	31 (24-42)	NS	-
Thrombomodulin	5764 (4919-6914)	6339 (5181-7117)	NS	-
Galectin-3	2654 (2323-3169)	2287 (2021-2649)	*	0.732
Collagen-IVα1	824 (557-1058)	650 (485-767)	NS	-
I-α-1/COL1A1	7261 (5124-11181)	11126 (8605-14162)	NS	-
Resistin	17358 (14048-27482)	9672 (7573-10533)	***	0.863

Table 7.6: Biomarker levels in streptococcal NSTIs with and without septic shock. Unit of measurement: pg/mL. Data are presented as median values with interquartile ranges (IQR). a Calculated after resampling as described in methods. b Mann-Whitney U test p-values after Benjamini-Hochberg adjustment: * ≤0.05; ** ≤0.01; *** ≤0.005. NS: non-significant. c AUC: Area under receiver operating curve. AUC values are presented exclusively for biomarkers with statistically significant differences obtained after Benjamini-Hochberg adjustment. d RF: Random forest. No. of trees: 100.000. Split: 6. Repetitions: 100. Accuracy: 72.6%. Age and gender are clinical parameters included in the RF modelling. e Non-shock>Septic shock. f Unit of measurement: $\times 10^9$/L. Missing six in the GAS cohort. g Unit of measurement: mg/L. Missing five in the GAS cohort.

	Septic shock	Non-shock	Mann-Whitney U[a]	AUC[c]	RF[d]	
	n = 64	n = 38	p-value[b]		Gini index	p-value
Interleukins						
IL-1α	60 (53-70)	46 (39-50)	***	0.826	1.47	0.2
IL-1β	20 (14-34)	12 (7-17)	***	0.736	0.94	1
IL-2	1209 (822-1538)	583 (411-1017)	***	0.784	1.13	0.88
IL-4	250 (222-286)	182 (152-213)	***	0.832	2.4	<0.01
IL-6	2676 (586-37,017)	216 (86-558)	***	0.829	2.72	<0.01
IL-10	91 (58-166)	44 (18-67)	***	0.786	1.24	0.14
IL-12p70	218 (181-292)	131 (75-196)	***	0.776	1.21	0.72
IL-13	1562 (1303-1793)	1218 (955-1474)	***	0.739	0.8	1
IL-17A	42 (20-90)	19 (12-30)	***	0.731	1.25	0.91
IL-18	566 (397-1106)	546 (291-856)	NS	-	1.27	0.91
IL-22	127 (116-151)	101 (87-121)	***	0.777	1.03	0.98
IL-23	2726 (486-8356)	1460 (196-4961)	NS	-	0.85	1
IL-36β/IL-1F8	19 (17-23)	15 (12-16)	***	0.847	2.28	<0.01
Chemokines						
CCL-2/MCP-1	1325 (754-2914)	387 (211-594)	***	0.847	2.79	0.02
CCL-4/MIP-1β	951 (937-1040)	767 (699-835)	***	0.813	1.82	0.09
CCL-5/RANTES	6056 (2155-14,518)	7490 (3684-12,088)	NS	-	1.17	0.98
CXCL-8/IL-8	74 (38-493)	19 (12-31)	***	0.806	2.42	0.04
CXCL-10/IP-10	19,426 (1298-385,626)	981 (371-11,054)	***	0.726	1.24	0.88
Adhesion molecules						
E-Selectin	156,013 (118576-214,019)	121,227 (71708-185,216)	NS	-	1	1
ICAM-1	737,901 (568,783-916,275)	616,286 (527,608-771,348)	NS	-	0.69	1
VCAM-1	49.9×10^5 (37.2×10^5 -70.5×10^5)	39.4×10^5 (21.3×10^5 -52.9×10^5)	NS	-	0.93	1
Matrix metalloproteases						
MMP-1	1961 (1366-4374)	1243 (856-1875)	*	0.682	0.91	1
MMP-8	45,030 (25367-73,789)	14,578 (8403-35,893)	***	0.73	1.2	0.92
MMP-9	4210 (2628-8410)	15,353 (5114-30,679)	***[e]	0.748[e]	1.64	0.38
Growth factors						
G-CSF	7544 (1989-60,986)	441 (269-1099)	***	0.858	3.82	0.02
Other						
C5/C5a	21,255 (13315-32,528)	22,401 (15613-34,755)	NS	-	1.46	0.77
Pentraxin-3	20,872 (10,810-30,902)	5735 (3163-12,991)	***	0.837	3.24	<0.01
TNFα	47 (29-72)	22 (17-30)	***	0.812	1.92	0.09
S100A8	1329 (939-2143)	707 (490-865)	***	0.82	3.85	<0.01
S100A9	3124 (1549-5714)	2151 (1477-3395)	NS	-	0.98	1
MPO	47,213 (36,078-61,287)	31,046 (24,733-40,880)	***	0.732	1.61	0.4
Fas-Ligand	41 (31-70)	30 (20-38)	**	0.719	1.7	0.31
Thrombomodulin	13,813 (10,269-16,996)	11,424 (7955-13,081)	*	0.678	1.6	0.5
Galectin-3	3564 (2973-4189)	3058 (2606-3503)	NS	-	0.76	1
Collagen-IVα1	3101 (2105-4597)	1846 (1365-2905)	***	0.736	1.09	0.99
I-α-1/COL1A1	10,861 (7826-23,185)	8030 (5302-13,384)	NS	-	1	1
Resistin	57,395 (45,002-67,308)	46,847 (31,006-60,136)	NS	-	0.63	1
Leucocytes[f]	11.9 (7.2-19.4)	16.5 (11.9-20.8)	NS	-	0.97	1
CRP[g]	284 (149-377)	270 (200-337)	NS	-	1.13	0.99

Table 7.7: Correlation between plasma biomarkers and disease severity assessed by Sequential Organ Failure Assessment (SOFA) score in streptococcal NSTI cases. Spearman correlation (r_s/ρ), significant at the 0.01 level (2-tailed) a SOFA score at admission. Three cases missing. b Missing six. c Missing five.

	SOFA score[a]	
	Correlation (ρ)	**p-value**
Interleukins		
IL-1α	0.509	<0.001
IL-1β	0.318	0.001
IL-2	0.435	<0.001
IL-4	0.472	<0.001
IL-6	0.494	<0.001
IL-10	0.556	<0.001
IL-12p70	0.454	<0.001
IL-13	0.369	<0.001
IL-17A	0.422	<0.001
IL-18	0.396	<0.001
IL-22	0.488	<0.001
IL-23	0.061	0.55
IL-36β/IL-1F8	0.502	<0.001
Chemokines		
CCL-2/MCP-1	0.507	<0.001
CCL-4/MIP-1β	0.481	<0.001
CCL-5/RANTES	-0.328	<0.001
CXCL-8/IL-8	0.557	<0.001
CXCL-10/IP-10	0.347	<0.001
Adhesion molecules		
E-Selectin	0.116	0.254
ICAM-1	0.342	<0.001
VCAM-1	0.355	<0.001
Matrix metalloproteases		
MMP-1	0.487	<0.001
MMP-8	0.515	<0.001
MMP-9	-0.482	<0.001
Growth factors		
G-CSF	0.437	<0.001
Other		
C5/C5a	-0.123	0.225
Pentraxin-3	0.358	<0.001
TNFα	0.536	<0.001
S100A8	0.493	<0.001
S100A9	0.146	0.151
MPO	0.488	<0.001
Fas-Ligand	0.267	0.008
Thrombomodulin	0.549	<0.001
Galectin-3	0.44	<0.001
Collagen-IVα1	0.414	<0.001
I-α-1/COL1A1	0.349	<0.001
Resistin	0.321	0.001
Leucocytes[b]	-0.14	0.18
CRP[c]	0.136	0.19

Table 7.8: Biomarker levels in streptococcal NSTIs by outcome. Unit of measurement: pg/mL. Data are presented as median values with interquartile ranges (IQR). a Severe outcome: Death or amputation within 90 days. Only cases with NSTIs affecting extremities were included. Those amputated before arrival at the study hospital were excluded. Amputation is defined as the surgical removal of all or part of a limb or extremity. b Calculated after resampling as described in methods section. c Mann-Whitney U test p-values after Benjamini-Hochberg adjustment: * ≤0.05; ** ≤0.01; *** ≤0.005. NS: non-significant. d AUC: Area under receiver operating curve. AUC values are presented exclusively for biomarkers with statistically significant differences obtained after Benjamini-Hochberg adjustment or RF modelling. e RF: Random forest. No. of trees: 100.000. Split: 6. Repetitions: 100. Accuracy: 79.7%. Age, gender and septic shock, hyperbaric oxygen treatment (HBOT) and intravenous immunoglobulin (IVIG) treatment are clinical parameters included in the RF modelling. f Unit of measurement: $\times 10^9$/L. Missing three. g Unit of measurement: mg/L. Missing two.

	Severe outcome[a]	Non-severe outcome	Mann-Whitney U[b]	AUC[d]	RF[e]	
	n = 16	n = 65	p-value[c]		Gini index	p-value
Interleukins						
IL-1α	65 (54-70)	53 (46-60)	NS	-	0.55	0.1
IL-1β	26 (15-40)	17 (11-24)	NS	-	0.28	0.96
IL-2	1099 (741-1441)	1038 (561-1253)	NS	-	0.33	0.85
IL-4	231 (217-286)	220 (179-261)	NS	-	0.29	0.94
IL-6	9885 (2599-66,661)	506 (168-1947)	NS	0.795	0.97	0.04
IL-10	127 (62-312)	63 (40-90)	NS	0.702	0.94	0.03
IL-12p70	250 (206-329)	182 (127-231)	NS	-	0.52	0.31
IL-13	1575 (1282-1837)	1409 (1154-1626)	NS	-	0.3	0.91
IL-17A	33 (17-167)	27 (17-61)	NS	-	0.51	0.37
IL-18	561 (387-1473)	593 (386-800)	NS	-	0.29	0.94
IL-22	134 (112-196)	120 (101-134)	NS	-	0.52	0.28
IL-23	2972 (101-6935)	1886 (519-8376)	NS	-	0.3	0.96
IL-36β/IL-1F8	19 (16-23)	17 (15-20)	NS	-	0.26	0.96
Chemokines						
CCL-2/MCP-1	988 (612-2556)	775 (343-1791)	NS	-	0.4	0.66
CCL-4/MIP-1β	913 (837-1033)	835 (760-955)	NS	-	0.34	0.76
CCL-5/RANTES	2906 (1430-10,603)	6861 (2892-12,481)	NS	-	0.84	0.1
CXCL-8/IL-8	230 (64-643)	33 (16-81)	NS	-	0.61	0.26
CXCL-10/IP-10	4355 (893-222,004)	6878 (824-383,300)	NS	-	0.35	0.74
Adhesion molecules						
E-Selectin	120,734 (80,256-193,336)	143,440 (103,933-187,719)	NS	-	0.44	0.6
ICAM-1	748,585 (507,738-1,002,047)	654,278 (544,169-863,507)	NS	-	0.56	0.31
VCAM-1	5.2×10^6 (4.6×10^6 -7.7×10^6)	4.1×10^6 (3.2×10^6 -6.3×10^6)	NS	-	0.47	0.58
Matrix metalloproteases						
MMP-1	2518 (1772-7568)	1405 (937-2736)	NS	-	0.38	0.79
MMP-8	51,096 (23,720-75,962)	24,620 (11,130-52,946)	NS	-	0.66	0.26
MMP-9	3217 (2320-10,830)	5701 (3844-21,384)	NS	-	0.43	0.68
Growth factors						
G-CSF	26,117 (3279-421,222)	1465 (430-8386)	NS	0.747	0.97	0.03
Other						
C5/C5a	17,937 (12,831-2247)	22,579 (14,177-37,416)	NS	-	0.52	0.51
Pentraxin-3	20,679 (11,514-25,231)	12,991 (5330-27,959)	NS	-	0.3	0.98
TNFα	36 (28-84)	29 (21-48)	NS	-	0.3	0.92
S100A8	1849 (1020-2720)	915 (703-1341)	NS	-	0.57	0.19
S100A9	1795 (1084-4787)	2616 (1477-4213)	NS	-	0.65	0.23
MPO	55,629 (38,269-80,820)	37,231 (29,011-53,639)	NS	-	0.6	0.32
Fas-Ligand	34 (25-87)	38 (27-52)	NS	-	0.68	0.25
Thrombomodulin	13,813 (10,301-18,267)	12,407 (9102-15,802)	NS	-	0.28	0.94
Galectin-3	3580 (3259-4123)	2370 (2628-3664)	NS	-	0.32	0.94
Collagen-IVα1	4131 (2243-6203)	2347 (1554-3300)	NS	0.734	1.37	<0.01
I-α-1/COL1A1	9341 (6926-19,013)	9130 (7160-17,516)	NS	-	0.83	0.17
Resistin	57,395 (49,644-67,740)	54,208 (31,893-63,171)	NS	-	0.29	0.97
Leucocytes[f]	9.3 (6.6-23.7)	13.9 (7.7-19.3)	NS	-	0.55	0.4
CRP[g]	168 (110-309)	271 (192-364)	NS	-	0.46	0.66

7.3.4 Mediator levels by streptococcal etiology

There were no significant differences in mediator levels in NSTI caused by GAS compared to SD in univariate analysis. According to the RF model, however, three markers higher in GAS cases (CXCL-10/IP-10, E-Selectin, and S100A9) differentiated the two etiologies, all with AUC >0.70 (Table 7.9). Restricting analysis to cases with septic shock generated similar results (Table 7.10).

7.3.5 Identification of biomarker profiles related to clinical categories using unsupervised hierarchical clustering

To explore further the differential inflammatory profiles of the NSTI and cellulitis cohorts, we performed an unsupervised hierarchical cluster analysis, including septic shock, severe outcome, and streptococcal etiology of NSTIs. Four main clusters were detected, as shown in Figure 7.1. The clusters with high or intermediate mediator levels (denoted cluster 1a and 1b, respectively) included 87% (20/23) of the cases with severe outcome (P=0.002), and 85% (57/67) of the cases with septic shock (P<0.001). Cluster 1a comprised GAS cases only. Sixteen of the 19 non-operated cellulitis cases were grouped in cluster 2, showing, in general, low levels of mediators.

7.3.6 Network and connectivity analysis

Network connectivity analysis was applied to assess interactions and to identify key response nodes among the mediators. Differences in connectivity were evident within the GAS cohort, whereas few significant connections were detected in the SD cohort (Table 7.11). In the mediator-mediator association networks, a similar pattern was seen. Several associations among mediators were retrieved for the GAS cohort, with distinctive connections and strong power of the connections, of which the connection between IL-17A and S100A9 in septic shock cases was most evident. In contrast, associations were generally weak in the SD cohort (Figure 7.2).

Biomarker responses after in vitro stimulation

In vitro stimulation of PBMC and human umbilical vein endothelial cells (HUVEC) with GAS and SD was performed to corroborate the host responses measured in plasma (Figure 7.3). Overall GAS triggered a higher IFN-γ, TNFα, and IL-1β response in PBMC as compared to SD. Similarly, when exposing HUVECs to the stimulated PBMC supernatants, the highest release of CXCL-10/IP-10 and E-Selectin responses were generally observed following stimulation with supernatants of PBMCs exposed to GAS. In contrast, stimulation with heat-killed SD bacteria induced a higher IL-1β and E-Selectin response as compared to heat-killed GAS, but not for S100A9, which was higher in GAS.

Table 7.9: Biomarker levels in monomicrobial NSTIs caused by either GAS or SD. Abbreviations: GAS: *S. pyogenes* (Group A streptococcus), SD: *S. dysgalactiae*. Unit of measurement: pg/mL. Data are presented as median values with interquartile ranges (IQR). a Calculated after resampling as described in methods. b Mann-Whitney U test p-values after Benjamini-Hochberg adjustment: * ≤0.05; ** ≤0.01; *** ≤0.005. NS: non-significant. c AUC: Area under receiver operating curve. AUC values are presented exclusively for biomarkers with statistically significant differences obtained after Benjamini- Hochberg adjustment or RF modelling. d RF: Random forest. No. of trees: 100.000. Split: 6. Repetitions: 100. Accuracy: 81.5%. Age, gender and septic shock are clinical parameters included in the RF modelling. e Unit of measurement: $\times 10^9$/L. Missing six in the GAS cohort. f Unit of measurement: mg/L. Missing five in the GAS cohort.

	GAS	**SD**	**Mann-Whitney U[a]**	**AUC[c]**	**RF[d]**	
	n = 88	n = 14	**p-value[b]**		**Gini index**	**p-value**
Interleukins						
IL-1α	55 (46-66)	49 (46-60)	NS	-	0.98	0.8
IL-1β	17 (12-29)	18 (9-29)	NS	-	1.14	0.69
IL-2	1018 (557-1395)	758 (585-1211)	NS	-	1.22	0.63
IL-4	227 (183-270)	223 (184-231)	NS	-	0.94	0.89
IL-6	702 (184-3570)	4603 (413-20,540)	NS	-	1.43	0.29
IL-10	74 (39-135)	60 (44-69)	NS	-	1.2	0.66
IL-12p70	199 (129-273)	183 (121-246)	NS	-	0.91	0.96
IL-13	1489 (1172-1729)	1282 (1016-1498)	NS	-	0.85	1
IL-17A	31 (17-74)	18 (11-30)	NS	-	1.17	0.81
IL-18	594 (384-1031)	423 (344-585)	NS	-	1.13	0.89
IL-22	123 (100-144)	112 (102-117)	NS	-	1.04	0.85
IL-23	2316 (525-8612)	508 (15-2521)	NS	-	1.68	0.4
IL-36β/IL-1F8	18 (15-21)	16 (15-19)	NS	-	1	0.76
Chemokines						
CCL-2/MCP-1	917 (405-1962)	671 (181-2070)	NS	-	1.29	0.47
CCL-4/MIP-1β	868 (761-1007)	890 (765-914)	NS	-	1.76	0.06
CCL-5/RANTES	6766 (2918-15,495)	2516 (1162-9006)	NS	-	1.18	0.81
CXCL-8/IL-8	38 (18-202)	101 (40-683)	NS	-	1.56	0.21
CXCL-10/IP-10	9616 (784-383,497)	893 (149-1508)	NS	0.726	4.19	0.01
Adhesion molecules						
E-Selectin	150,022 (116,004-212,739)	93,763 (61,876-108,329)	NS	0.742	2.18	0.04
ICAM-1	692,105 (569,574-883,187)	507,738 (395,507-833,182)	NS	-	2.13	0.06
VCAM-1	4.2×10^6 (3.3×10^6 -6.3×10^6)	4.6×10^6 (3.5×10^6 -6.3×10^6)	NS	-	0.74	1
Matrix metalloproteases						
MMP-1	1566 (1067-3167)	1637 (665-2437)	NS	-	1.08	0.85
MMP-8	33,640 (12481-65,414)	22,312 (9384-38,223)	NS	-	1	0.93
MMP-9	5608 (3349-15,353)	4420 (1446-23,091)	NS	-	2.28	0.11
Growth factors						
G-CSF	1830 (431-16,675)	11,025 (892-39,393)	NS	-	0.93	0.96
Other						
C5/C5a	23,522 (15,547-34,128)	14,136 (11,511-21,060)	NS	-	2.76	0.08
Pentraxin-3	17,157 (6336-28,517)	11,873 (6549-19,059)	NS	-	0.95	0.99
TNFα	35 (22-61)	29 (26-37)	NS	-	0.96	0.96
S100A8	1014 (647-1803)	1037 (711-1758)	NS	-	1.02	0.83
S100A9	3059 (1685-5450)	1075 (671-1776)	NS	0.834	2.71	0.02
MPO	42,537 (32,105-57,588)	34,100 (27,343-50,148)	NS	-	1.38	0.47
Fas-Ligand	36 (27-55)	32 (20-41)	NS	-	1.18	0.75
Thrombomodulin	12,972 (8845-16,336)	12,014 (9217-13,659)	NS	-	1.11	0.91
Galectin-3	3318 (2770-4011)	3259 (2888-3647)	NS	-	0.93	0.98
Collagen-IVα1	2604 (1823-4188)	2051 (1599-3353)	NS	-	0.67	1
I-α-1/COL1A1	10,052 (7496-23,772)	6863 (3586-14,472)	NS	-	1.43	0.61
Resistin	55,283 (42,589-65,419)	48,427 (31,788-67,177)	NS	-	1.19	0.81
Leucocytes[e]	14.2 (8.2-20.4)	15.7 (6.6-23.7)	NS	-	1.53	0.31
CRP[f]	296 (200-368)	148 (66-261)	NS	-	2.44	0.08

Table 7.10: Biomarker levels in septic shock cases of GAS and SD aetiology. Abbreviations: GAS: *S. pyogenes* (Group A streptococcus), SD: *S. dysgalactiae*. Unit of measurement: pg/mL. Data are presented as median values with interquartile ranges (IQR). a Calculated after resampling as described in methods. b Mann-Whitney U test p-values after Benjamini-Hochberg adjustment: *≤0.05; **≤0.01; ***≤0.005, NS: non-significant. c AUC: Area under receiver operating curve. AUC is presented exclusively for biomarkers with statistically significant differences obtained after Benjamini-Hochberg adjustment or RF modelling. d RF: Random forest. No. of trees: 103. Split: 6. Repetitions: 100. Accuracy: 80.3%. Age and gender are clinical parameters included in the RF modelling.

	GAS	SD	Mann-Whitney U[a]	AUC[c]	RF[d]	
	n = 55	n = 9	p-value[b]		Gini index	p-value
Interleukins						
IL-1α	60 (54-72)	56 (46-61)	NS	-	0.2	0.92
IL-1β	19 (14-34)	20 (16-30)	NS	-	0.16	1
IL-2	1224 (939-1595)	825 (663-1228)	NS	-	0.28	0.71
IL-4	259 (222-291)	231 (228-252)	NS	-	0.19	0.97
IL-6	1756 (472-32280)	7815 (6495-57752)	NS	-	0.52	0.12
IL-10	91 (65-238)	61 (50-98)	NS	-	0.3	0.55
IL-12p70	218 (176-301)	206 (184-254)	NS	-	0.22	0.91
IL-13	1577 (1355-1829)	1307 (1214-1573)	NS	-	0.21	0.97
IL-17A	60 (23-101)	29 (17-36)	NS	-	0.31	0.71
IL-18	594 (404-1229)	438 (378-671)	NS	-	0.24	0.81
IL-22	134 (117-161)	116 (110-122)	NS	-	0.32	0.5
IL-23	3139 (549-8612)	1613 (260-2967)	NS	-	0.16	0.99
IL-36β/IL-1F8	20 (18-24)	18 (16-19)	NS	-	0.21	0.92
Chemokines						
CCL-2/MCP-1	1328 (799-3148)	979 (643-2806)	NS	-	0.22	0.94
CCL-4/MIP-1β	973 (842-1067)	901 (810-921)	NS	-	0.31	0.51
CCL-5/RANTES	6861 (2642-16285)	2321 (1162-2700)	NS	-	0.76	0.1
CXCL-8/IL-8	70 (31-475)	430 (67-697)	NS	-	0.42	0.26
CXCL-10/IP-10	44816 (3076-385730)	959 (454-1441)	NS	0.784	1.78	0.02
Adhesion molecules						
E-Selectin	159518 (133984-222069)	96073 (61893-166475)	NS	0.711	1.33	0.01
ICAM-1	783535 (623014-986502)	525519 (452138-649200)	NS	-	0.54	0.09
VCAM-1	5.1×10^6 (3.7×10^6-6.9×10^6)	4.8×10^6 (4.4×10^6-7.0×10^6)	NS	-	0.13	1
Matrix metalloproteases						
MMP-1	1801 (1366-5269)	2089 (1936-2583)	NS	-	0.23	0.91
MMP-8	50037 (27086-73857)	28636 (9584-38223)	NS	-	0.32	0.57
MMP-9	4517 (2738-8009)	2619 (1402-10661)	NS	-	0.57	0.35
Growth factors						
G-CSF	6505 (1830-54462)	29158 (16048-64437)	NS	-	0.43	0.29
Other						
C5/C5a	23775 (14829-33031)	12780 (5880-17105)	NS	-	0.95	0.06
Pentraxin-3	22623 (11696-34215)	12304 (10723-19059)	NS	-	0.19	0.99
TNFα	55 (31-74)	29 (28-37)	NS	-	0.23	0.91
S100A8	1316 (939-2234)	1522 (1150-1940)	NS	-	0.22	0.91
S100A9	3400 (1685-6327)	1109 (540-1776)	NS	0.824	1.44	0.03
MPO	48221 (39156-61922)	36641 (34074-52898)	NS	-	0.41	0.47
Fas-Ligand	44 (31-70)	33 (31-41)	NS	-	0.31	0.64
Thrombomodulin	14422 (10473-16996)	12811 (9217-14117)	NS	-	0.18	0.96
Galectin-3	3587 (2992-4311)	3403 (2888-3647)	NS	-	0.34	0.6
Collagen-IVα1	3040 (2132-4688)	3267 (1861-3745)	NS	-	0.17	0.99
I-α-1/COL1A1	12659 (8025-48404)	9125 (5102-15266)	NS	0.663	1.08	0.04
Resistin	57915 (48677-66258)	50885 (41773-67177)	NS	-	0.2	0.95
Leucocytes	11.7 (7.4-20.1)	16.3 (3.6-17.7)	NS	-	0.38	0.81
CRP	302.5 (194-383)	113.0 (60-166)	NS	-	1	0.06

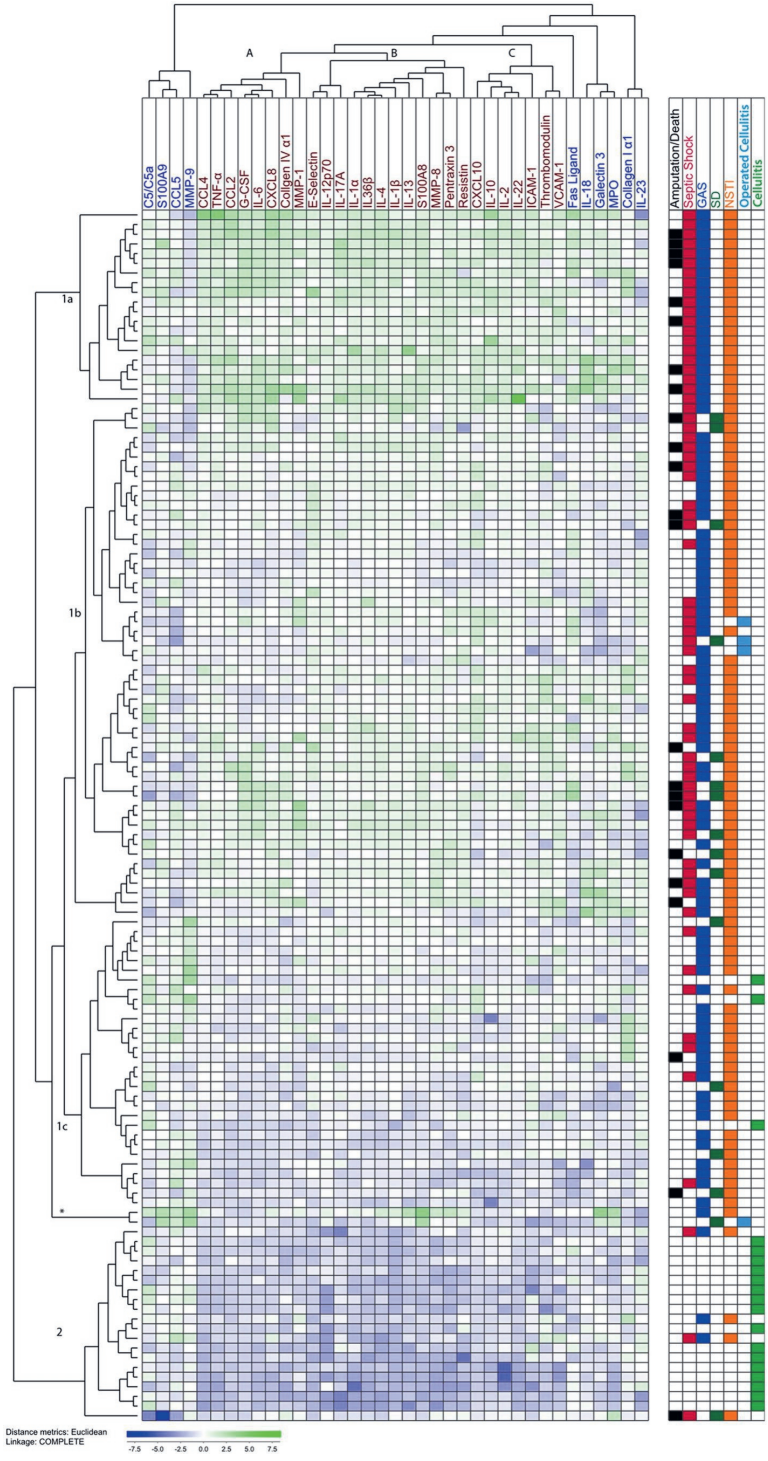

A
B
C
C5/C5a
S100A9
CCL5
MMP-9
CCL4
TNF-α
CCL2
G-CSF
IL-6
CXCL8
Collagen IV α1
MMP-1
E-Selectin
IL-12p70
IL-17A
IL-1α
IL36β
IL-4
IL-1β
IL-13
S100A8
MMP-8
Pentraxin 3
Resistin
CXCL10
IL-10
IL-2
IL-22
ICAM-1
Thrombomodulin
VCAM-1
Fas Ligand
IL-18
Galectin 3
MPO
Collagen I α1
IL-23
Amputation/Death
Septic Shock
GAS
SD
NSTI
Operated Cellulitis
Cellulitis
1a
1b
1c
*
2
Distance metrics: Euclidean
Linkage: COMPLETE
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5

Figure 7.1

Table 7.11: Connectivity analysis. Samples were aggregated by microbe (GAS vs SD). The differential connectivity was calculated for the comparison of septic shock development for GAS and SD. Abbreviations: GAS: *S. pyogenes* (Group A streptococcus), SD: *S. dysgalactiae*.

Mediators	GAS				SD			
	Connectivity		Differential connectivity	p-value	Connectivity		Differential connectivity	p-value
	SS	No SS			SS	no SS		
Interleukins								
IL-1α	0.71	0.66	0.05	0.334	0.26	0.28	0.02	0.894
IL-1β	0.85	0.58	0.27	0.002	0.18	0.21	0.03	0.831
IL-2	1	0.78	0.22	0.009	0.25	0.32	0.07	0.586
IL-4	0.58	0.79	0.21	0.008	0.32	0.4	0.08	0.709
IL-6	0.9	0.25	0.65	0.001	0.48	0.19	0.29	0.029
IL-10	0.23	0.28	0.05	0.203	0.28	0.18	0.1	0.492
IL-12p70	0.53	0.57	0.04	0.438	0.12	0.18	0.06	0.63
IL-13	0.8	0.87	0.07	0.171	0.27	0.3	0.03	0.807
IL-17A	0.43	0.42	0.01	0.722	0.24	0.28	0.04	0.814
IL-18	0.94	0.6	0.34	0.001	0.3	0.21	0.09	0.697
IL-22	0.34	0.59	0.25	0.003	0.18	0.38	0.2	0.23
IL-23	0.7	0.35	0.35	0.001	0.15	0	0.15	0.153
IL-36β/IL-1F8	0.79	0.94	0.15	0.03	0.33	0.32	0.01	0.94
Chemokines								
CCL-2/MCP-1	0.73	0.44	0.29	0.001	0.42	0.15	0.27	0.031
CCL-4/MIP-1β	0.48	0.76	0.28	0.002	0.44	0.29	0.16	0.347
CCL-5/RANTES	0.83	0.57	0.26	0.001	0.32	0	0.32	0.01
CXCL-8/IL-8	0.85	0.4	0.45	0.001	0.31	0.19	0.12	0.274
CXCL-10/IP-10	1	0.46	0.54	0.001	0.06	0	0.06	0.657
Adhesion molecules								
E-Selectin	0.49	0.57	0.08	0.118	0.08	0.18	0.1	0.473
ICAM-1	0.81	0.36	0.45	0.001	0.26	0.24	0.02	0.88
VCAM-1	0.89	0.35	0.54	0.001	0.5	0.28	0.22	0.12
Matrix metalloproteases								
MMP-1	0.56	0.54	0.02	0.574	0.4	0.15	0.25	0.03
MMP-8	0.76	0.47	0.29	0.001	0.18	0.14	0.04	0.749
MMP-9	0.38	0.22	0.16	0.004	0.31	0.09	0.22	0.073
Growth factors								
G-CSF	0.82	0.54	0.28	0.001	0.26	0.26	<0.01	0.971
Other								
C5/C5a	0.13	0.28	0.15	0.023	0.21	0.2	0.01	0.906
Pentraxin-3	0.99	0.36	0.63	0.001	0.18	0.3	0.12	0.337
TNFα	0.48	0.22	0.26	0.001	0.32	0.27	0.05	0.729
S100A8	0.97	0.47	0.5	0.001	0.33	0.09	0.24	0.059
S100A9	0.56	0.38	0.18	0.002	0.46	0.06	0.4	0.001
MPO	0.72	0.76	0.04	0.401	0.13	0.15	0.02	0.882
Fas-Ligand	0.43	0.52	0.09	0.095	0.24	0.31	0.07	0.653
Thrombomodulin	1.16	0.54	0.61	0.001	0.32	0.16	0.16	0.194
Galectin-3	0.63	0.43	0.2	0.003	0.17	0.9	0.08	0.456
Collagen-IVα1	0.59	0.42	0.17	0.013	0.35	0.15	0.2	0.098
I-α-1/COL1A1	0.71	0.39	0.32	0.001	0.41	0.27	0.14	0.301
Resistin	0.66	0.45	0.21	0.013	0.22	0.3	0.08	0.526

Figure 7.1: Unsupervised hierarchical clustering analysis of plasma levels of 37 mediators in 102 NSTI patients and 23 cellulitis patients with streptococcal aetiology. Euclidean distance and complete linkage were applied in the clustering analysis. The dendrogram on the top (biomarkers) and on the left side (cases) of the heat map form clusters. The threshold is set to midpoint of the longest branch. Cluster 1a is made by cases with generally high levels of biomarkers, while cluster 1b represent cases with intermediately high values of biomarkers, cluster 1c represent neutral/low values and cluster 2 represents low levels of biomarkers. * constitutes a group of two cases, situated between cluster 1a and 1b/1c with respect to biomarker levels.

7.4 Discussion

This study demonstrates a profound immune activation in NSTI caused by β-hemolytic streptococci and the way it differs from non-necrotizing infections with the same pathogens. We have identified several single mediators and broader host response profiles associated to type of infections, streptococcal species, disease severity and outcome. The findings illustrate that different immunological pathways and pathogenic processes are involved or predominate along the spectrum of moderate and severe streptococcal SSTIs. The different patterns observed in our study can be exploited to advance the development of much needed new diagnostic tools to aid clinical decisions and management.

In NSTI, early recognition, surgical debridement and appropriate antimicrobial therapy contribute to reduce mortality and improve patient outcomes **(Goh et al. 2014; Nawijn et al. 2020)**. It is therefore a great concern that misdiagnosing at admission is frequent **(Goh et al. 2014)**. Use of scorings systems like the LRINEC score **(Wong, Khin, et al. 2004)**, based on routine laboratory values, has turned out to be of limited value as adjuncts to clinical judgement **(Fernando et al. 2019; Hsiao et al. 2020)**. In our study, however, several inflammatory mediators and mediator profiles showed promising diagnostic and prognostic accuracy, outperforming both CRP and leucocytes.

Data on use of biomarkers in the diagnosis of NSTIs are scarce **(Saccenti and Svensson 2020; Medina et al. 2021; Ling et al. 2023)**. Hansen et al. **(Hansen, Rasmussen, Svensson, et al. 2017)** studied cytokine responses in NSTIs, observing higher levels of IL-6 and TNFα in streptococcal compared to other NSTIs. A contemporary retrospective study, revealed IL-6 as the most accurate cytokine to distinguish NSTIs from severe SSTIs **(Ling et al. 2023)**. We have previously explored the pathogenesis in NSTIs in the INFECT cohort identifying differential host-pathogen-interactions by aetiology **(Thänert et al. 2019; Jahagirdar, L. Morris, et al. 2022)** in **Chapter 5** itnameref5. Recently, in the same cohort we identified that Thrombomodulin was a promising general marker for NSTI, whereas G-CSF, S100A8 and IL-6 were associated to septic shock **(Medina et al. 2021)** in **Chapter 6** *The Race Against Time: Discriminatory Plasma Biomarkers*. The present study is restricted to GAS and SD aetiology, and biomarker candidates that may be used for separating streptococcal NSTIs from streptococcal cellulitis are unravelled. Based on data herein, biomarker profiles (including e.g. IL-1β, CXCL-8/IL-8, TNFα and possibly Thrombomodulin) could become valuable adjunctive tools to rapidly decipher severe (i.e. NSTI) from

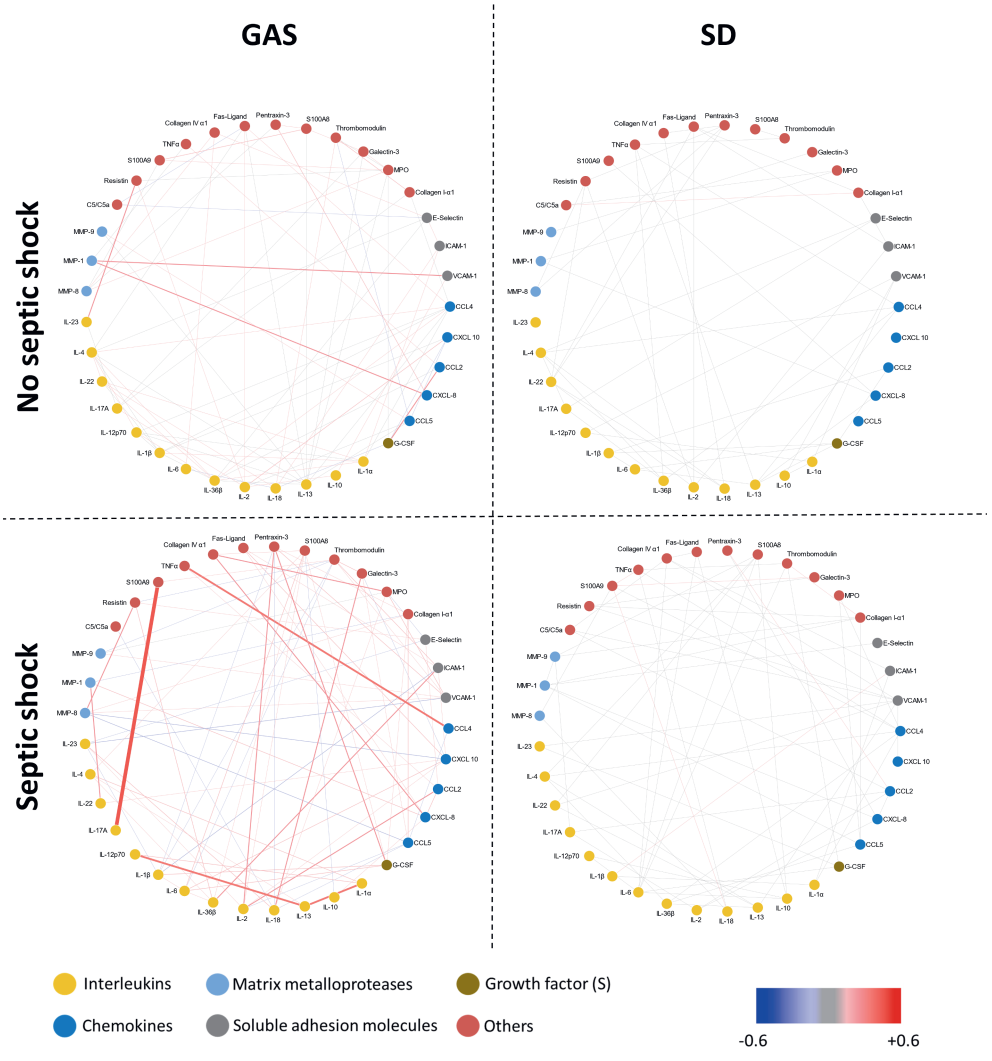


Figure 7.2: Association networks of *S. pyogenes* (group A streptococcus; GAS) and S. dysgalactiae (SD) NSTIs with and without septic shock. Mediators are categorized by colors. Each mediator (protein) is represented by a node, and the connection between nodes are represented by links (or edges), and the strength of the links are weighted by the thickness of the connection.

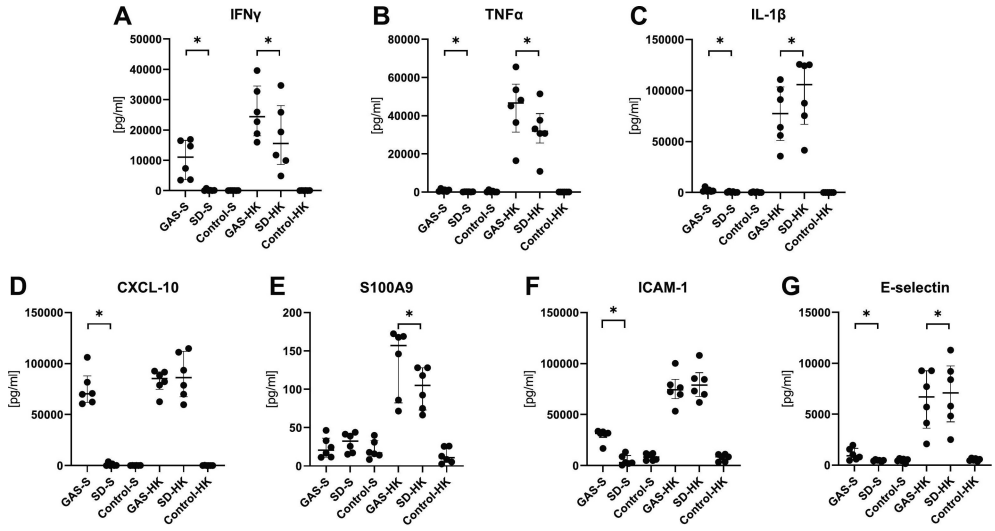


Figure 7.3: Cell responses during in vitro stimulation with GAS and SD. Concentrations of selected analytes in cell culture media after stimulation of PBMC (A-C) or HUVEC (D-E). Isolated PBMCs were stimulated with bacterial supernatant (S) or heat-killed bacteria (HK) for 24 h. The supernatants of the PBMC stimulations were then used for stimulation of HUVEC cells for another 24 h. Bacterial culture media and PBS were used as negative controls for supernatant and heat-killed bacteria stimulations, respectively. Stars indicate significance (p-value<0.05).

less severe (i.e. cellulitis) infections. Although Thrombomodulin was not one of the mediators reaching significance in the RF model, as opposed to our former study, it displayed a very high AUC (0.91) and a low p-value (P = 0.059).

Distinguishing GAS from SD NSTIs based on clinical findings, is difficult, and of limited clinical importance. Of the three mediators displaying discriminant ability in our study, CXCL-10/IP-10 appear as a central chemokine in GAS NSTIs, as previously published **(Medina et al. 2021; Jahagirdar, L. Morris, et al. 2022; Thänert et al. 2019)**, and merits further investigation.

Biomarker profiles may also be useful in risk stratification and prognostic evaluation. In the hierarchical cluster analysis, we found that 87% of the patients with a severe outcome, and 85% of those with septic shock clustered in two groups with similar profiles. However, the method is descriptive, complex and not feasible to apply in every-day-clinical practice. Nevertheless, hierarchical cluster analysis may contribute to creation of novel hypotheses. Multiplex profiling or use of combinations of biomarkers offers advantages compared to single biomarkers, as it can portray the concomitant pro- and anti-inflammatory pattern expressed by the patient **(Wright et al. 2021; B. M. Tang et al. 2010)**. The powerful pro-inflammatory response in streptococcal NSTI, especially when eliciting STSS, is probably a main reason for mortality **(Gottlieb et al. 2018; Siemens, Snäll, et al. 2020)**. In two studies of NSTIs of all etiologies, both pro-inflammatory (IL-1β, IL-6, G-CSF and TNFα) and anti-inflammatory (IL-10) cytokines were associated with severity, mortality or amputation **(Hansen, Rasmussen, Svensson, et al. 2017; Hedetoft, Garred, et al.**

2021). Bulger and coworkers showed that higher levels of plasma chemokines and cytokines (TNFα, IL-6, CXCL-8/IL-8, and CCL-2/MCP-1) were correlated to poorer clinical outcome **(Bulger et al. 2018)**. In our study, we found that IL-6, along with IL-10, G-CSF and Collagen-IVα1 could predict severe outcome in streptococcal NSTI patients, all four displaying biomarker potential (AUC >0.70).

In streptococcal disease, bacterial toxins, but also immune cells and other host factors contribute to tissue damage and systemic inflammatory reactions **(Siemens, Snäll, et al. 2020; Johansson, Thulin, et al. 2010; Norrby-Teglund, Chatellier, et al. 2000)**. In GAS NSTI, STSS is frequent and > 60% in our GAS NSTIs cohort had septic shock **(Bruun, Rath, et al. 2021; Low 2013)**. Key mediators of STSS are the superantigens **(Norrby-Teglund, Thulin, et al. 2001; Commons et al. 2014)**, which activate T cells in an unconventional manner resulting in a massive cytokine response, including release of IL-6, CXCL-8/IL-8, and CCL-2/MCP-1, which all were independent markers of shock in our study **(Siemens, Snäll, et al. 2020; Proft et al. 2022)**. At-present, 13 superantigens have been identified in GAS **(Siemens, Snäll, et al. 2020; Commons et al. 2014)**. In contrast, SpeG is currently the only superantigen gene identified in a substantial number of SD isolates **(Commons et al. 2014)**, and its activity and involvement in toxic shock is unclear. Although the profound immune activation observed in the present study may reflect a major role for superantigens in the systemic response to streptococcal NSTIs, it is unclear to what extent the differences observed between GAS and SD infections are due to superantigen activity. Previous findings have demonstrated that the local cytokine response in severe SSTIs caused by GAS resembles that of a systemic superantigen induced response **(Norrby-Teglund, Thulin, et al. 2001)**, but the role for superantigens at the tissue level in less severe infections, like cellulitis, is not clear.

Notably, IL-1β levels were elevated in NSTIs compared to cellulitis. This pro-inflammatory cytokine has been inferred as a key mediator facilitating GAS NSTIs **(Siemens, Snäll, et al. 2020; Chella Krishnan et al. 2016)**.

Network analyses revealed strong associations between several mediators in NSTIs caused by GAS. The connections may reflect the predominant role of certain virulence factors, including superantigens, in activating pathways. In contrast, the absence of strong connectivity in SD NSTI networks may reflect lack of superantigen activity. Also, it could reflect the diversity of this population, where comorbidities and age may have a greater impact on the host responses observed. Of note, this result may be influenced by the low number of SD cases. Nevertheless, the in vitro stimulation experiments suggest that also SD has the ability to induce a broad and strong activation of the immune system.

Together with S100A8, S100A9 constitutes the heterodimer calprotectin, a well-known damage-associated molecular patterns (DAMPs) molecule, highly concentrated in e.g. phagocytes **(Vogl, Eisenblätter, et al. 2014)**. It has been shown that IL-17A can induce S100A8 and S100A9 expression in keratinocytes **(S. C. Liang et al. 2006)**. Furthermore, the complex S100A8/100A9 is a known endogenous activator of Toll-like receptor 4, subsequently promoting endotoxin-induced septic shock **(Vogl, Tenbrock, et al. 2007)**, and has been associated with increased risk of mortality in septic shock patients **(Dubois, Marcé, et al. 2019)**. In our study, the powerful association of S100A9 and IL-17A in GAS septic shock cases suggests that the same pathway may also be involved in gram-positive sepsis.

A main limitation of this study is the low number of severe cellulitis cases in the control group, but resampling was applied to overcome the uneven numbers of patients in the different groups. Moreover, the profound differences seen between NSTI and cellulitis cases would mitigate the under-powered situation of the comparison. Collection of plasma was done in a standardized fashion, but only at the study hospitals, and not at admittance to the primary hospital (for the NSTI cohort). However, for the referred patients, median time from admission at primary hospital to admission at study hospital was not more than14 and 18 h, (for SD and GAS cases, respectively) **(Bruun, Rath, et al. 2021)**.

Major strengths of this study include a predefined study protocol, and the multicenter prospective patient enrollment, contributing to inclusion of a homogenous patient cohort. This cohort of cases caused by GAS and SD is the largest to date. In addition, this study included comparable control cohorts. The statistical methods applied decreased the likelihood of committing type 1 errors. These strengths made it possible to identify robust associations pointing at specific immunological pathways and biomarkers in SSTIs of streptococcal etiology. The study therefore also adds to the identification of candidate targets for personalized therapy in streptococcal NSTI.

In summary, this prospective study of streptococcal NSTIs compared with cellulitis cases, identified systemic inflammatory mediators significantly associated to type of infection as well as severity and prognosis, both single mediators and wider profiles. The study also highlights interactions and patterns of immune activation that may direct search for future targets for therapy in NSTIs.

7.5 Author contribution

A.N.T. is project coordinator of the INFECT study. E.R., T.B. and S.S. conceived the biomarker study. O.H., S.S. and M.N. are national investigators and have contributed to study design and coordinated study conduct. M.B.M. is responsible for the database and contributed to patient inclusion and data collection, as did E.R., T.B., S.S., O.O., T.N., N.H., M.N. and O.H. E.R., L.M.P.M., S.J., T.B., S.S., K.A.M., J.K.D., E.S., V.A.P.M.d.S., M.S., and A.N.T. contributed to analysis or interpretation of data or both. E.R. drafted the publication. All authors contributed to the writing and approved the final version.

7.6 Acknowledgements

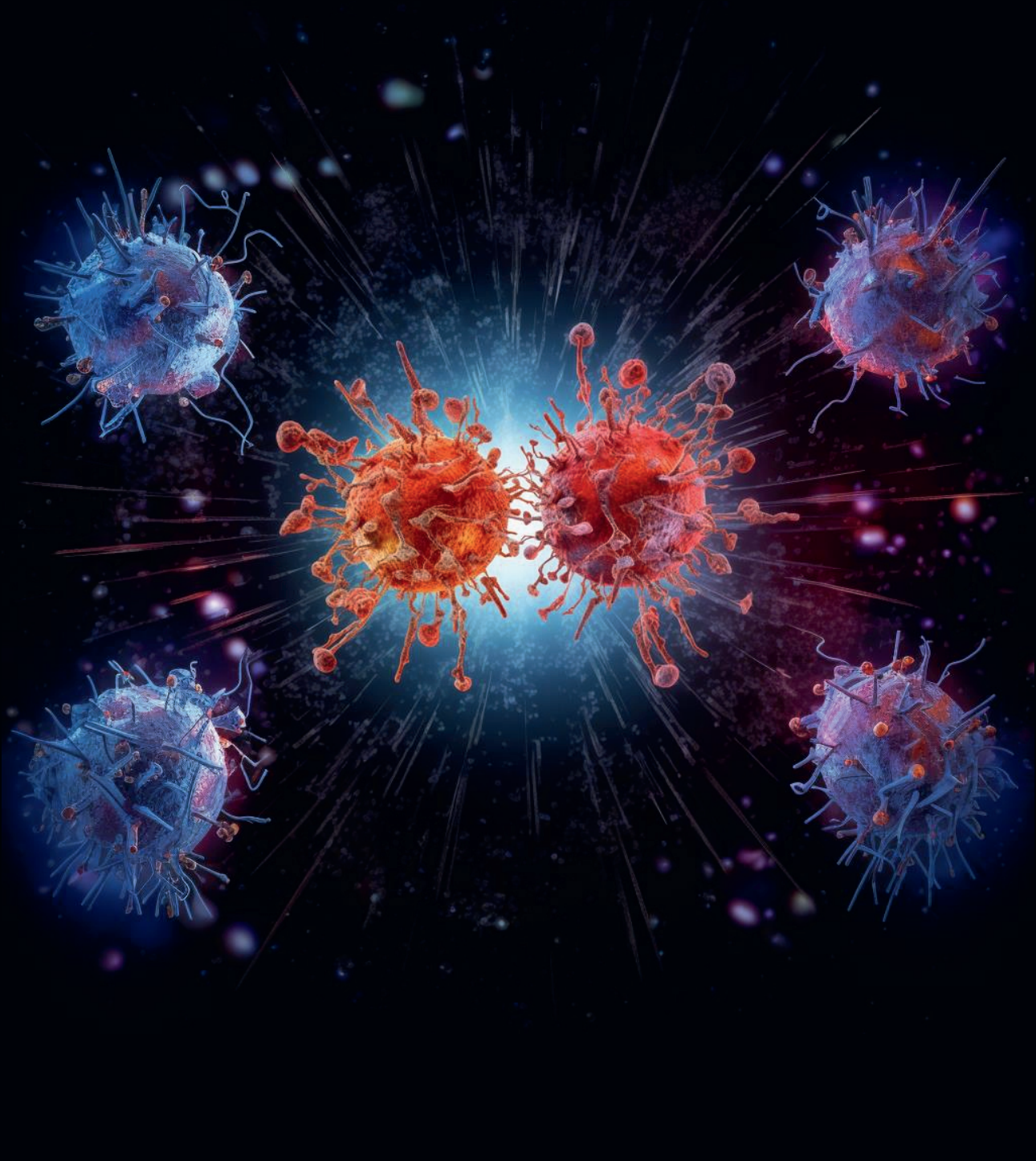
INFECT study group: Oddvar Oppegaard, Haukeland University Hospital, Bergen, Norway; Torbjørn Nedrebø, Haukeland University Hospital, Bergen, Norway; Morten Hedetoft, Department of Anaesthesia, Hyperbaric Unit, University Hospital Rigshospitalet, Copenhagen; Michael Nekludov, Department of Anaesthesia, Surgical Services and Intensive Care, Karolinska Institute, Karolinska University Hospital, Stockholm, Sweden.

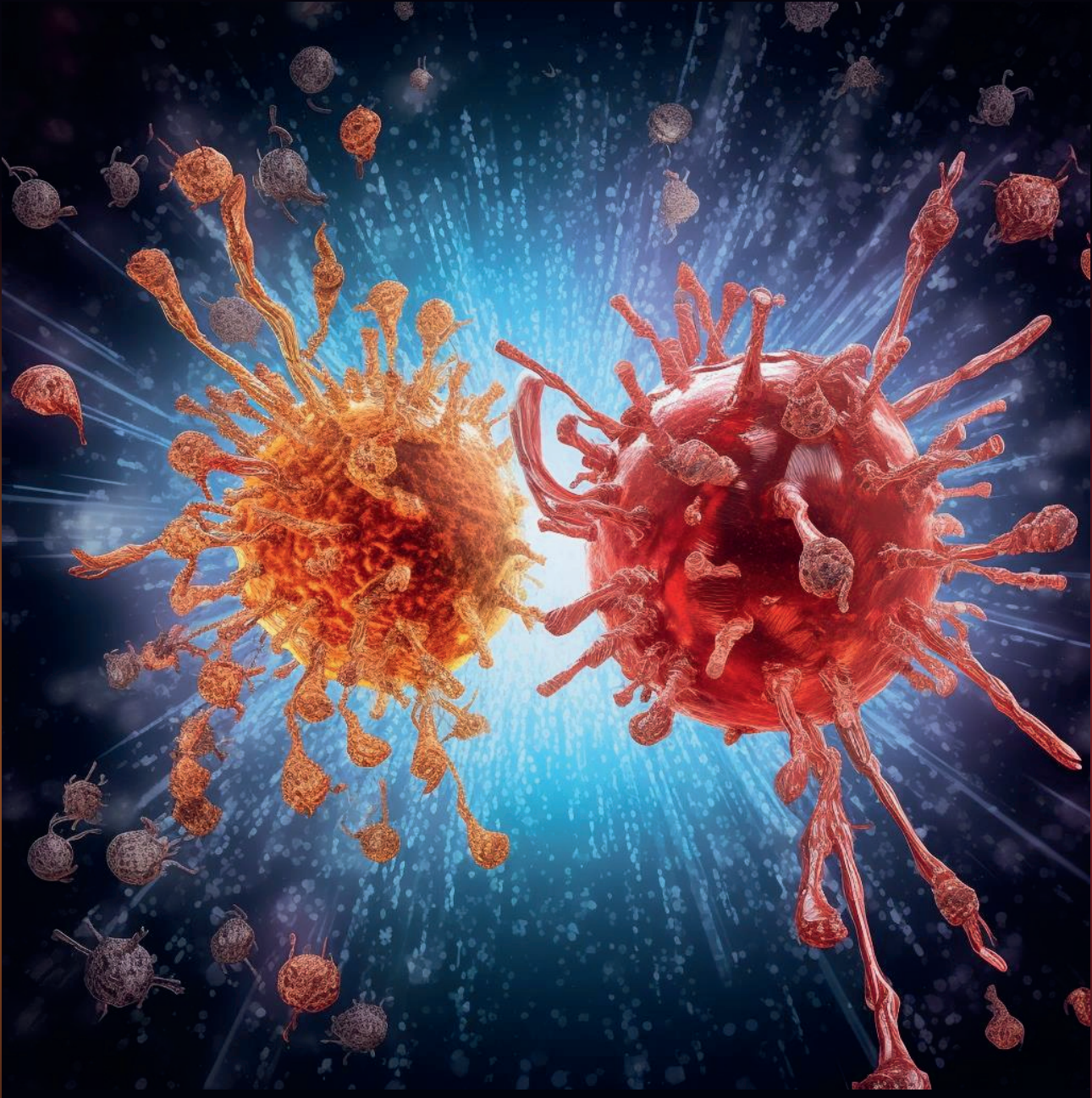
We would like to acknowledge the expert technical assistance of Kristin Rye, and Karen M. Hagen for excellent assistance with the ELISA analyses. Øystein Bruserud is gratefully acknowledged for valuable discussions. In addition, thanks are due to all co-workers of the INFECT project, to patients and relatives for participation in

Publication

this study, and to the microbiological departments at every study hospital for routine- and species diagnostics.

Chapter 8





Sanjeevan Jahagirdar[1], Merijn Lamers[1*], Chen Kuang[1,18$*], Oriol Basallo Clariana[1,19$*], Laura M. Palma Medina[4], Knut Anders Mosevoll[2,3], Eivind Rath[2], Trond Bruun[2,3], Mattias Svensson[4], Ole Hyldegaard[5,6], Anna Norrby-Teglund[4], Steinar Skrede[2,3], INFECT Study group[#], Vitor A. P. Martins dos Santos[7,8], Edoardo Saccenti[1]

***Contributed equally**

$Current affiliations. Work done when affiliated with Wageningen University & Research

#INFECT study group (Trond Bruun, Eivind Rath, Torbjørn Nedrebø, Per Arnell, Anders Rosen, Morten Hedetoft, Martin B. Madsen, Mattias Svensson, Johanna Snäll, Ylva Karlsson, & Michael Nekludov)

Turn to page 377 for author affiliations

This chapter is prepared for publication

Deadly Dance: Understanding the Interplay of Pro- and Anti-Inflammatory Cytokines

Abstract

Introduction: Necrotising soft tissue infections (NSTI) are aggressive, life-threatening bacterial infections that are associated with a high risk of mortality and a severe decline in quality of life. The extensive damage to the soft tissue is not solely attributable to the activity of the micro-organisms. Immune cells, along with host-derived factors such as cytokines and other signalling molecules, also contribute to the resulting hyper-inflammation. These cytokines are frequently characterised as pro- or anti-inflammatory based on their specific impact on the state of inflammation.

Methods: This study is an exploratory discourse to determine whether the pairwise relationships among cytokines, plasma analytes, and analogous proteins offer valuable insights for discriminating NSTI cases. Ratios of the concentrations and parameters derived from a cytokine dynamic model (Baker model), fitted to the concentrations, serve as proxies for these relationships. We employ XGBoost models to evaluate the information content encapsulated in these ratios and parameters, compared to the standalone concentrations.

Results: The discriminatory capacity of cytokine, plasma analyte, and analogous protein concentration ratios and parameters exhibited a marginally superior ability to differentiate (a) NSTI from controls, (b) microbial aetiology, and (c) patient outcomes, in comparison to standalone concentrations. Additionally, particular ratios and parameters demonstrating substantial capacity to distinguish subtypes. The ratio CCL-4/MIP-1β $\Longleftrightarrow$ Thrombomodulin had an extremely good ability to differentiate between NSTI and the pragmatic classification of non-NSTI with the model classifying patients with a ratio less than 0.1 as NSTI

Conclusion: By generating data-driven hypotheses focused on the pairwise relationships between cytokines and other plasma analytes involved in the human immune response, we delineate ratios and parameters that embody the pro- and anti-inflammatory characteristics of cytokines. The findings enrich our understanding of the fundamental mechanisms implicated in NSTI, offering potential avenues for devising more efficacious diagnostic and prognostic approaches for this condition.

8.1 Introduction

Necrotising soft tissue infections (NSTI) are rapidly progressing devastating bacterial infections with a high risk of mortality characterised by impairment and injury in any layer of the soft tissue compartment **(Hua et al. 2022; D. L. Stevens and Bryant 2017)**. These rare infections are aggressive in nature causing severe loss in quality of life often due to extensive tissue loss and amputations during treatment **(Urbina et al. 2021; Suijker et al. 2020; M. B. Madsen, Bergsten, et al. 2020)**. NSTIs have often been broadly classified into polymicrobial (Type I) caused by many microorganisms and monomicrobial (Type II) caused by a single microorganism.

Although the comprehension of the complete mechanism involved in the progression of the disease is incomplete, it has been shown that the impairment is a result of both the action of microorganisms as well as the immune cells and host-derived factors causing hyper-inflammation **(Siemens, Snäll, et al. 2020; Medina et al. 2021)**. Cytokines and other signalling molecules have been shown to play a role in this inflammation **(Thänert et al. 2019)**. Cytokines and other signaling molecules play a crucial role in inflammation during NSTI and the release of pro-inflammatory cytokines has been associated with the occurrence of septic shock **(Hansen, Simonsen, et al. 2015)**. Similarly, IL-1 signialing has been found to be associated with host-protective functions in streptococcal NSTI **(Richter et al. 2021)**. We study the ratios and parameters derived from the concentrations of chemokines, soluble adhesion molecules, interleukins, matrix metalloproteases, and other analytes in blood plasma (collectively referred as analytes in this study) of NSTI patients enrolled in the INFECT study **(Medina et al. 2021; Rath et al. 2023)** to unravel mechanistic insights. INFECT is the world's largest multicenter, prospective cohort study on NSTI patients **(M. B. Madsen, Skrede, et al. 2019)**.

Using computational experimentation for exploratory modeling has had a positive effect in many different fields dealing with complex systems and irreducible uncertainties **(Kwakkel et al. 2013)**. In traditional systems medicine, approaches are often divided into data-driven top-down analyses and model-driven bottom-up analyses **(R.-S. Wang et al. 2015)**. We take cues from both approaches and design an exploratory study to help improve our mechanistic understanding and generate hypotheses based on the pair-wise relationships of these analytes involved in the human immune response. These relationships and interactions between analytes are as important to understanding the complex systems as the concentrations of the analytes measured in the patients. Many studies have identified ratios of analytes rather than concentrations as fundamental biomarkers for detecting various illnesses **(Russell et al. 2019; Ye et al. 2020; J. C. Chan et al. 2017; Klanderman et al. 2019)**.

Analytes have often been described as having a pro-inflammatory or an anti-inflammatory function based on their role of promoting or reducing inflammation **(Dinarello 2000)**. In this study, we estimate parameters by optimising the Baker cytokine dynamic model **(M. Baker 2015)** such that the measured concentrations of analytes represents a steady state in the model. The parameters in the Baker model have been interpreted biologically based on the pro- and anti-inflammatory roles played by the analytes in the dynamic models.

In this study, we further evaluate the information content present in the pair-

wise relationships of the analytes (represented by ratios and model parameters) in comparison with the information contained in the analyte concentrations. We compare these by using the concentrations, ratios, and model parameters separately to train a machine-learning algorithm to classify patients based on known factors and then evaluate the information present in these derived data sets to classify the NSTI patients. An overview of the scheme used in this study is shown in figure 8.1.

We trained the classification models based on biologically relevant questions that have been pursued by the INFECT consortia in previous studies (**Chapter 6**, **Chapter 7**) to differentiate between (a) patients and various controls, (b) patient outcomes, and (c) microbial aetiology **(Medina et al. 2021; Rath et al. 2023)**. The classification models and the patients included in the comparisons are shown in figure 8.2. We use the feature importance coupled with biologically interpreted parameters to explore hypotheses pertaining to relevant biological questions.

8.2 Methods

8.2.1 Study design

This study is an exploratory discourse to ascertain if the interactions between cytokines, analytes and similar proteins hold important insights towards bacteremia in NSTI cases. The first approach in many such studies is often to analyse the ratios of the proteins as ratios have been shown to be good predictors of reaction rates **(A.-K. Petersen et al. 2012)**. In this study, we also attempt to use model parameters to express a variety of relationships that may be expressed between the analytes. Dynamic models in biology often take the form of

$$\frac{d}{dt}x = f\left(x, t, \mu, \beta\right), \tag{8.1}$$

where, x is generally a state vector like the concentration of i^{th} analyte (c_i), μ represent actuation parameters like the initial concentration $c(0)_i$ and β represents the biological parameters like reaction rates. In this study, we explore whether or not ratios and model parameters β can (a) hold information regarding the disease and (b) shed some light on mechanistic insights toward our understanding of NSTIs. The scheme of the design is shown in figure 8.1

8.2.2 Patient cohorts and cytokine data

The analyte concentrations are based on plasma samples collected from surgically confirmed necrotising soft tissue infection patients (n=251) that were enrolled in the multicenter INFECT study. These samples were from 5 hospitals namely, Blekingesjukhuset (Karlskrona, Sweden), Haukeland University Hospital, Karolinska University Hospital, Righospitalet (Copenhagen, Denmark), and Sahlgrenska University Hospital. Detailed information pertaining to the sample collection is reported in **(Medina et al. 2021)** *The Race Against Time: Discriminatory Plasma Biomarkers*, section and the patient characteristics and outcomes are reported in **(M. B. Madsen, Skrede, et al. 2019)**. Analyte concentrations from three different control groups were also analysed. These patients were classified as

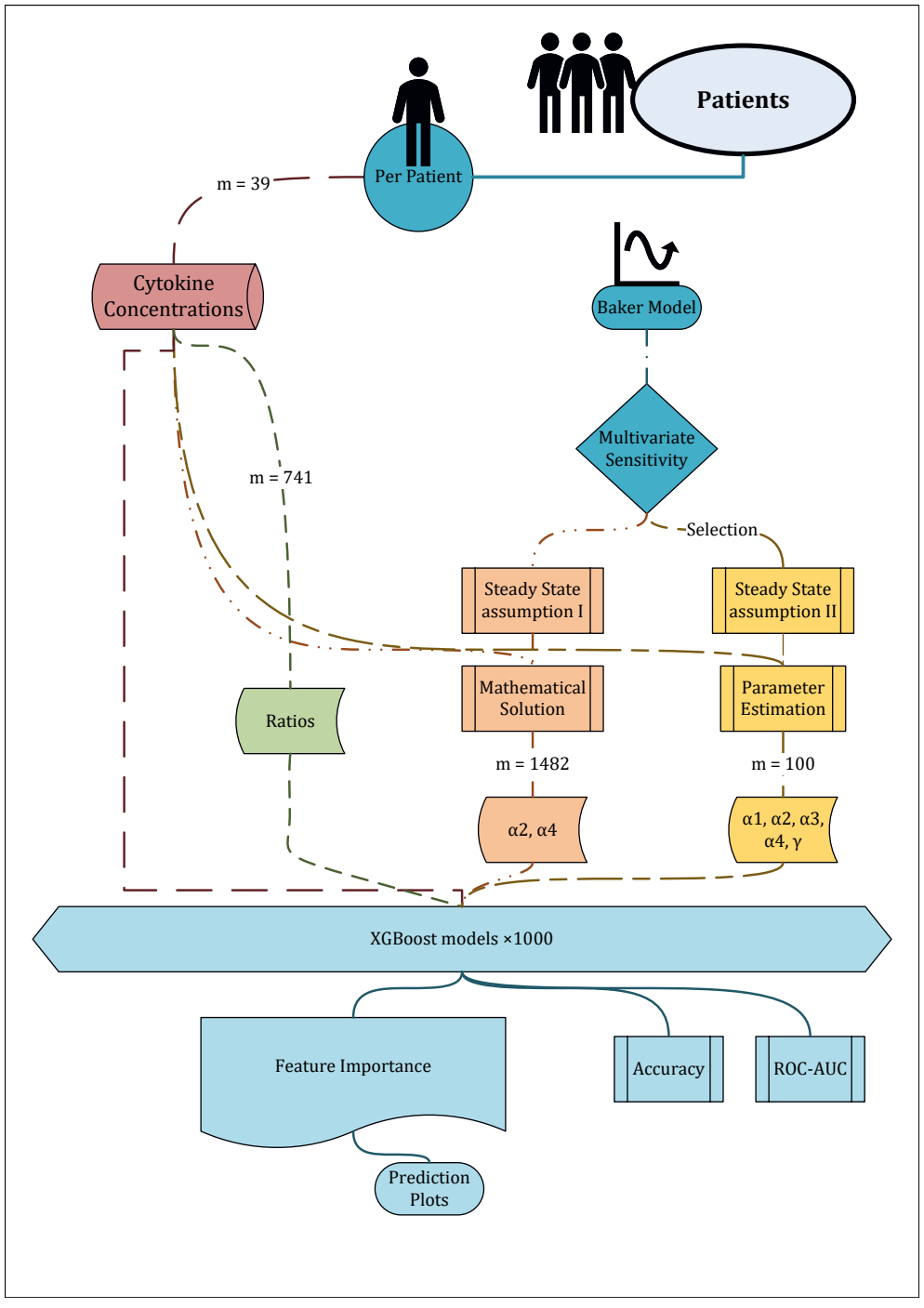

Patients
Per Patient
m = 39
Cytokine Concentrations
Baker Model
Multivariate Sensitivity
m = 741
Selection
Steady State assumption I
Steady State assumption II
Mathematical Solution
Parameter Estimation
Ratios
m = 1482
m = 100
α2, α4
α1, α2, α3, α4, γ
XGBoost models ×1000
Feature Importance
Accuracy
ROC-AUC
Prediction Plots

Figure 8.1

Figure 8.1: A flowchart of the general study design. The concentrations of 39 analytes collected are used to calculate the ratios and model parameters for every patient. XGboost models are then used to assess the informational contents in the ratios and model parameters in comparison with the analyte concentration values. The XGboost models were trained for two-group classification problems addressing biological questions of interest explored in previous studies. All the classifications are shown in figure 8.2. The feature importance of these models is then studied to gain mechanistic understanding and generate hypotheses.

- Surgical controls (n=20) who underwent elective surgeries for non-infectious reasons
- Cellulitis patients (n=19) who suffer from bacterial infections with similar symptoms but less severe outcomes and
- Non-NSTI patients (n=20) who were suspected NSTI patients but were found to be non-necrotic upon surgery.

In total 39 analyte concentrations were analysed that included chemokines, interleukins, adhesion molecules, matrix metalloproteases and others. Complete information on the analytes measured and the process of measurement can be found in **(Medina et al. 2021; Rath et al. 2023)** *The Race Against Time: Discriminatory Plasma Biomarkers*.

8.2.3 Data imputation

The analyte concentrations had missing values, mainly due to the values being outside of the measurement equipment range. A detailed imputation strategy was applied to the missing values of cytokine concentrations based on the information on account of the left-centerdness or right-centerdness of the missing values. Different imputation strategies were implemented if the out-of-range (OOR) values were greater than maximum or less than minimum. Detailed information can be found in **(Rath et al. 2023)** *The Immune System Responds: Systemic Immune Activation Profiles*.

8.2.4 Data transformation

Prior to fitting the cytokine dynamic model to the concentration data, the concentration values were scaled using max-scaling. This was done to avoid completely altering the model behavior based on the magnitudes of the concentrations of the analytes. We used max-scaling to constrain the range of every analyte between 0 and 1. The analytes were assumed to have a minimum possible concentration of 0 as a negative concentration is biologically infeasible. The scaling for the i^{th} analyte is given by

$$c_i^s = \frac{c_i}{max(c_i)}. \tag{8.2}$$

where c_i^s is the scaled concentration of the i^{th} analyte with constraints $[0, 1]$ and c_i is the original concentration of the i^{th} element.

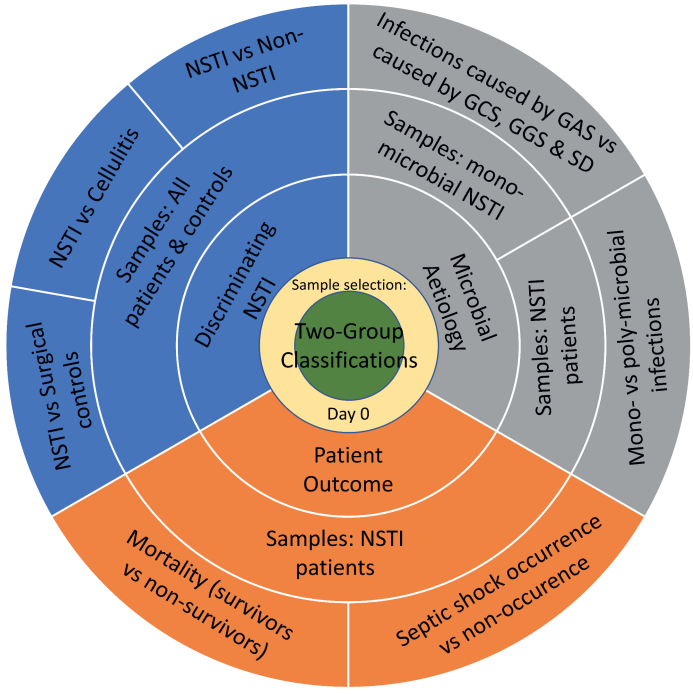


Figure 8.2: An overview of all the classifications the XGboost models were trained to predict. The various levels show both the biological organisation of the classification problems and the sample selection for training particular models. The blue, orange and grey colours categorise the biological questions pertaining to discriminating NSTI from controls, NSTI patient outcomes and causal microbial aetiology respectively. In sample selection, day 0 pertains to the measurement being taken on day of admission to the hospital. GAS references Group A Streptococcus as the causal micro-organism and SD references *Streptoccoccus dysgalactae* as the causal micro-organism. The models were trained on the 7 classification problems outlined in the outermost layer.

8.2.5 Ratios

Ratios of the concentrations of all the analytes were derived in a pair-wise fashion. The ratios of the i^{th} & j^{th} analyte was determined by

$$r_{ij} = \frac{c_i}{c_j}. \tag{8.3}$$

where r_{ij} is the ratio. Duplicates were considered to have the same information, i.e. the information gained from r_{ij} = the information gained from r_{ji} and thus all combinations of analytes were considered to calculate the ratios. Note: When estimating the parameters further in the study, all permutations were considered.

Table 8.1: All the parameters from the Baker model used in equations 8.4 & 8.5

Parameter	Representation
β_0	background pro-inflammatory production rate
β_1	maximum production rate for pro-inflammatory cytokine caused by pro-inflammatory binding
β_2	pro-inflammatory cytokine concentration at which $\phi(p)$ reaches half it's max value
β_3	efficiency of the inhibition by the anti-inflammatory cytokine
β_4	anti-inflammatory cytokine concentration at which $\theta(a)$ reaches half it's max value
β_5	maximum production rate for anti-inflammatory cytokine caused by pro-inflammatory binding
β_6	pro-inflammatory cytokine concentration at which $\psi(p)$ reaches half it's max value
m_1	positive integers representing co-operative binding to receptors of secreted cytokines
m_2	positive integers representing co-operative binding to receptors of secreted cytokines
m_3	positive integers representing co-operative binding to receptors of secreted cytokines

8.2.6 Baker model

Bacterial interactions and complex cross-talks with the human immune system can propel pro- and anti-inflammatory processes. In order to analyse parameters that model this type of pro- and anti-inflammatory processes, we chose an existing model that was designed to study the cytokine concentration in time and account for the pro- and anti-inflammatory dynamics that often govern these interactions. The Baker model **(M. Baker 2015)** postulates the dynamics of pro- and anti-inflammatory cytokines while retaining the interaction between the two cytokines. An additional rationale for using this model was that the model properties and resulting parameter spaces have been thoroughly explored **(W. Zhang, Jang, et al. 2019)**. The model incorporates two ordinary differential equations (ODEs) that take into account

- (a) degradation of the cytokines/analytes,
- (b) the positive ($\phi(p)$ and $\psi(p)$) feedback of pro-inflammatory behaviour and
- (c) the negative ($\theta(a)$) feedback of the anti-inflammatory behaviour.

These feedbacks are modelled as a hill function. The pro- and anti- inflammatory cytokine dynamics is modelled with

$$\frac{dp}{dt} = -d_p p + \phi(p)\theta(p), \tag{8.4a}$$

$$\frac{da}{dt} = -d_a a + \psi(p). \tag{8.4b}$$

where,

$$\phi(p) = \beta_0 + \beta_1 \frac{p^{m_1}}{\beta_2^{m_1}},\ \theta(a) = \beta_3 \frac{\beta_4^{m_2}}{\beta_4^{m2} + a^{m_2}},\ \psi(p) = \beta_5 \frac{p^{m_3}}{\beta_6^{m_3} + p^{m_3}}. \tag{8.5}$$

such that, p and a represent the concentration of the pro- and anti-inflammatory cytokines, i.e. at time t, $p = c(t)_i$ for the i^{th} cytokine. The representation of the parameters is given in table 8.1

Zhang et al. redefine the model to remove all the dimensions from the model and reduce the number of parameters to five parameters and 3 cooperative binding factors **(W. Zhang, Jang, et al. 2019)**. We use this non-dimensional model as well given our constraints on the available data. The concentration and time parameters

are made dimensionless by $p = p^* c_2$, $a = a^* c_4$ and $t = t^*/d_a$ where, p^*, a^* & t^* are the non-dimension equivalents. Thus the ODEs can be transformed as

$$\frac{dp}{dt} = -\gamma p + \frac{1}{1 + a^{m_2}} \left(\alpha_1 + \alpha_2 \frac{p^{m_1}}{1 + p^{m_1}} \right), \tag{8.6a}$$

$$\frac{da}{dt} = -a + \alpha_4 \frac{p^{m3}}{\alpha_3^{m_3} + p^{m_3}}. \tag{8.6b}$$

where,

$$\alpha_1 = \frac{\beta_0 \beta_3}{\beta_2 d_a},\ \alpha_2 = \frac{\beta_1 \beta_3}{\beta_2 d_a},\ \alpha_3 = \frac{\beta_6}{\beta_2},\ \alpha_4 = \frac{\beta_5}{\beta_4 d_a},\ \gamma = \frac{d_p}{d_a}. \tag{8.7}$$

The biological interpretations of these parameters are given in table 8.2

Table 8.2: The biological interpretations of the new parameters introduced in the dimensionless model from equations 8.6 & 8.7

Parameter	Biological Interpretation
γ	ratio of the rate of pro-inflammatory and anti-inflammatory decay
α_1	background production rate of the pro-inflammatory cytokine
α_2	maximum rate of pro-inflammatory cytokine production
$\alpha 3$	concentration of pro-inflammatory cytokine when anti-inflammatory production is half the maximum level
$\alpha 4$	maximum rate of anti-inflammatory cytokine production

8.2.7 Sensitivity of the parameters in the model

In order to understand the dynamics of the ODE model and determine the most influential parameters in the system, we perform a multi-variate sensitivity analysis. This allows us to study the effects of different parameters on the output of the model. We use the Monte Carlo simulation method to calculate the sensitivity. We further estimate the collinearity as a measure of the identifiability of all parameters and parameter sets. These results in ensemble were used to determine the parameters to estimate in 8.2.10

8.2.8 Assumptions

Construction of dynamic cytokine models taking into consideration the physico-chemistry can provide mechanistic insight into the functions of the immune system. However, there are both technical and ethical limitations in collecting dynamic data from blood samples of patients. This means that the analyte concentration data we possess represent two-time points separated by three days where as the analyte concentrations can fluctuate in the magnitude of hours or even minutes. Thus, we have to make certain assumptions in order to pursue this type of exploratory strategy. We make the following assumptions.

- We build a dimensionless model as discussed in the section *Baker model*. This is important as the time point of the measurements taken by the clinicians can vary from patient to patient and so can the time of prognosis and admission to

the hospital. As analyte levels can fluctuate in a matter of minutes and the measurements were taken on admission to the hospital and on day 3 of the patient being in the hospital, we can not account for a realistic time measurement.

- We make the assumption that cytokines either perform a pro- or anti-inflammatory function. This assumption played a role in the model selection as well. pro- and anti-inflammatory functions of analytes have been observed in biochemistry. However, we make a simplistic assumption that, when studying the pair-wise relationship between two analytes, one analyte takes a pro-inflammatory role and the other an anti-inflammatory role. This allows us to model two analytes i and j such that in one model the i^{th} analyte takes the pro-inflammatory role and the j^{th} analyte takes the anti-inflammatory role and in another model, the j^{th} analyte takes the pro-inflammatory role against the i^{th}. This situation is ofcourse distinct from biology where certain analytes are determined to be pro- and certain determined to be anti-inflammatory. Be that as it may the case, nuances persist and it is difficult to find consensus on the classification of all analytes into either pro- or anti-inflammatory categories furthermore this classification is convoluted by the fact that some analytes act as pro-inflammatory in some situations and as anti-inflammatory in others. Therefore, we found these mathematical permutations more impervious to future discoveries than the biological classifications of today.

- We make the assumption that the concentration of analytes measured in plasma represents a steady state. In biology, analytes are rarely in a steady state, however, studies have shown the benefits of making such an assumption both on the ease of modelling and the quality of the interpretations. The steady state assumption is discussed further in-depth in section *Assumptions*.

- In the section *Baker model* we discussed the parameters m_1, m_2 & m_3 as representing the co-operative binding to the analyte receptors. These parameters are positive integers. We follow the lead of **(W. Zhang, Jang, et al. 2019)** and make a constructive assumption to fix their values so that they act more as a constant than parameters.

$$m_1 = m_2 = m_3 = 2. \tag{8.8}$$

  We understand that each analyte may not have the same integer representing its cooperative binding, but it is unrealistic to find such information for all the cytokines and analytes in our study through literature research.

8.2.9 Steady state assumption

There are two plausible assumptions of steady state we made. In the first assumption, we propose the analyte concentration measurement taken on the day of arrival to the hospital (day 0) is the level that the analyte concentration remains steady forever. in

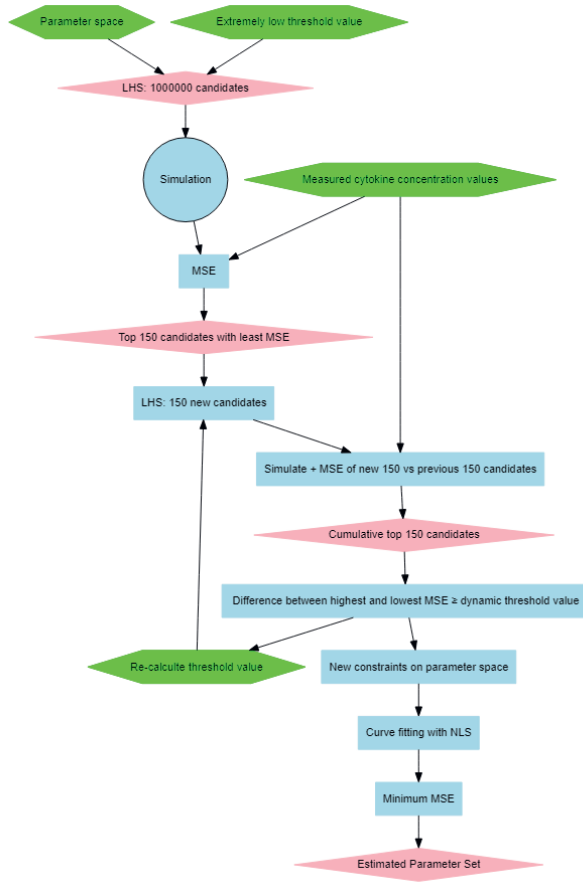


Figure 8.3: Overview of the algorithm used to estimate the parameters of the baker model. The algorithm follows a two-step parameter estimation approach that involves controlled random sampling and least-squares curve fitting.

the sense that,

$$\frac{dp}{dt} = 0 = -\gamma p + \frac{1}{1 + a^{m_2}}\left(\alpha_1 + \alpha_2 \frac{p^{m_1}}{1 + p^{m_1}}\right) \tag{8.9a}$$

$$\frac{da}{dt} = 0 = -a + \alpha_4 \frac{p^{m3}}{\alpha_3^{m_3} + p^{m_3}}. \tag{8.9b}$$

Taking equation 8.9 into account, equation 8.10 can be re-written as the following,

$$\alpha_2 = \frac{(p^2 + 1)((a^2 + 1)\gamma p - \alpha_1)}{p^2} \text{ with, } p \neq 0 \tag{8.10a}$$

$$\alpha_4 = a + \frac{\alpha_3^2 a}{p^2} \text{ with, } p \neq 0. \tag{8.10b}$$

In the second assumption, we assume that the analytes in healthy humans are at a very low concentration as can be found in the literature reported by various databases. The analyte concentration changes from the low level to the level measured when the patient suffering from NSTI is admitted to the hospital (day 0). The change in the concentration of analyte reaches a steady state at the measured value. Thus, $\frac{dp}{dt} = \frac{da}{dt} = 0$ at time = t. We set $t = 10000$ as a representative of a very large value as the model is dimensionless and time t does not have a unit.

8.2.10 Estimating parameters

Steady state assumption I

We explain here two parameter estimation strategies as we have explored two different definitions of steady state. In the first definition, we decide to explore the parameters α_2 & α_4. This decision is taken from the multi-variate sensitivity analysis that we perform in *Sensitivity of the parameters in the model* as seen from supplementary folder S3 Chapter 8. We solve the ODEs for α_2 & α_4 in equation 8.11. We use the parameter space exploration performed by Zhang et.al **(W. Zhang, Jang, et al. 2019)** to fix the values of $\gamma = 1.5$ and $\alpha_1 = 0.025$ in the first equation and the value of $\alpha_3 = 0.5$ in the second equation. This allows us to solve the equations for the values of α_2 & α_4

Steady state assumption II

In the second strategy, we assume the concentrations to reach a steady state at the measured levels. Here, we estimate all the parameters in the models except for the co-operative binding parameters that were fixed in 8.8. We realise that estimating multiple parameters in this setting would lead to numerous well-fit solutions to the equations. However, in this study, we are more interested to explore the parameters as proxy for the pair-wise relationships between the analytes than accurately predicting the exact dynamics occurring in the human body. Thus our strategy focused on pre-defining a multi-dimensional parameter space and finding the best ranking parameter sets in this solution space. To rank and find the best parameter set that fits all the assumptions we follow a two step parameter estimation approach. The entire strategy is summarised in 8.3

Controlled random sampling

In the first step, we perform a controlled random sampling on the parameter space. We kick-start the process by randomly sampling a million parameter values using the Latin hypercube sampling (LHS) method **(W.-L. Loh 1996)**. Every parameter set is plugged into the model and simulated. A mean square error (MSE) is calculated with respect to the actual measured values. MSE is calculated for the i^{th} analyte measurement as

$$MSE = \frac{\sum \left(c_{sim}(SS) - c_i(\text{day0 measurement})\right)^2}{n}. \tag{8.11}$$

where, c_{sim} is the simulated concentration at time $t = SS$. SS is the time when the system has reached a steady state. c_i is the actual concentration measurement taken

in hospital for the i^{th} analyte. Once we calculate the MSE for every parameter set, we rank the parameter sets based on their associated MSE value. The lower the MSE value, the higher the rank of the parameter set. The top 150 parameter sets also referred to as candidates are selected and the percent difference expressed in decimal form is calculated. This difference is compared to a pre-determined threshold and if the difference is less than the threshold, we sample 150 new candidates using LHS and the entire process repeats. The process stops when the difference between the 1st and 150th ranked parameter sets is less than that of the threshold. One of the issues with the process described above is that it is difficult to estimate the computational time required to perform it and it has the potential to be stuck in an endless loop with minimal information gain. To solve for this uncertainty, we established a dynamic threshold where we initiate a sliding threshold on a sigmoid curve with respect to the time elapsed. This way we allocate an exact computational time to the process. The dynamic threshold curve is shown in 8.4 and by the equation

$$\lambda = \frac{1}{1 + (\frac{x}{1-x})^{-B}} \text{ where, } x = \frac{\text{time elapsed}}{\text{total time}}. \tag{8.12}$$

Curve-ftting

The second step of the two-step parameter estimation strategy was to define new constraints on the parameter space based on the top 150 candidates from the previous step and then use a nonlinear solver to horn in on the so-called global minima of the MSE. We utilised Matlab's least-quares nonlinear solver to minimise the MSE function. The parameter set received from this method was chosen as the estimated parameter set.

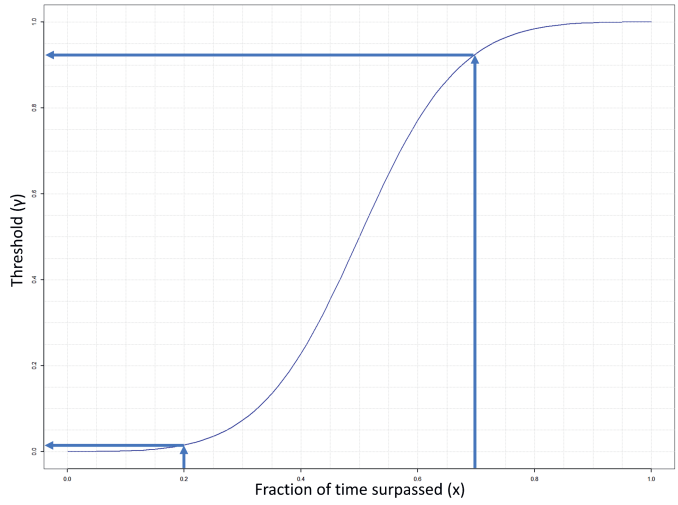


Figure 8.4: A graph depicting the threshold curve that was implemented to regulate the controlled random sampling process.

This parameter estimation needed to be performed on every analyte pair (1432) for every patient (310). To reduce the computational task, we pre-selected the analyte pairs based on the ability of the parameters determined in *Steady state assumption I* to discriminate NSTI patients using the area under the the receiver operating characteristic curve (AUC-ROC). The chosen analyte pairs and all the AUC values are shown insupplementarty file S3 Chapter 8.

8.2.11 ROC curve

Receiver Operating Characteristic (ROC) curves are graphical plots between sensitivity (also known as True positive rate) and 1-specificity (also known as false positive rate or 1-true negetive rate). We used the estimated value of the parameter as a classifier to classify between certain patient outcomes and infection aetiology and build the ROC curves as the threshold is varied over the range of the parameter values. The area under the curve (AUC) is taken of this ROC curve to measure the performance of the parameter values to classify these patient outcomes. The AUC can be interpreted as the probability that the parameter value will rank the classification of the correct patient outcome better than the randomly chosen incorrect one **(Fawcett 2006)**.

8.2.12 Assessing the information content of the parameter values

We define the information content of the parameter values by their ability to effectively distinguish between different patient outcomes, characteristics, and infection aetiology. A common way to establish this ability is by using the performance metrics of a model. In this approach, a model is trained on different data sets and classification problems. The performance of the different models gives us information about the information contained in a data set while performing the classification in comparison to the other data sets. This can be translated as the pair-wise relationships containing information regarding classifying NSTI sub-groups if the model performances are equivalent or better than when the models are built using concentration values alone. We used the Xgboost algorithm to train using the data on all the classification problems. Methods based on performing a sequence of repetitive binary partitions of a data set and then inferring predictions from the ends of the terminal split (also known as decision trees) have proven to be hugely successful in the past two decades. Xgboost also takes advantage of such an ensemble of decision trees and has been shown to be a very fast and accurate predictor especially when dealing with very large and complicated datasets **(T. Chen and Guestrin 2016; T. Chen, T. He, et al. 2015)**.

We train the Xgboost model to perform predictions on two-group classifications. The models were built with a max decision tree depth of s splits and a learning rate of 0.1. s was adjusted based on the number of variables (m) in the data that was used.

$$\text{s} = 4;\ \text{cytokine concentrations};\ m = 39, \tag{8.13a}$$

$$\text{s} = 16;\ \text{ratios \& unbiased ratios};\ m = 741, \tag{8.13b}$$

$$\text{s} = 32;\ \alpha_2 \ \&\ \alpha_4 \text{ from Steady State I};\ m = 1482, \tag{8.13c}$$

$$\text{s} = 8;\ \text{select parameters from Steady State II};\ m = 100. \tag{8.13d}$$

Samples were given weights based on the ratio of the number of samples in each classification group in each data set.

$$\text{scaled weight} = \frac{n_{\text{total}} - n_{\text{Classification G1}}}{n_{\text{Classification G1}}}. \tag{8.14}$$

where, $n_{\text{Classification G1}}$ is the number of samples in the training data that are assigned the 1^{st} classification group. We calculate the AUC-ROC to evaluate the model performance. With the addition of every decision tree in the model, we calculate the AUC

of the model both on training and testing datasets. The model stops adding decision trees to the ensemble as soon as the AUC on the test dataset decreases irrespective of the increase in the training data to prevent overfitting. Once there had been 50 decision tree additions with no increase in the AUC on test data, the algorithm was made to stop and the final model was the model with all the decision trees before the AUC on test data decreased. The entire process of sampling and building the model was repeated 1000 times in order to account for sampling biases and the best, means and the 95% confidence intervals of the evaluation parameters were reported. The feature importances of the models were studied and reported with respect to the measure gain. We further evaluate the top features based on their SHAP values to determine the effect the feature values have on the model classifications. SHAP values, an acronym for SHapley Additive exPlanations, were used for explaining the Xgboost models and understanding the contribution of each input feature to the prediction by estimating the expected change in the model prediction when the feature is observed, taking into account all possible combinations of other features **(Lundberg and S.-I. Lee 2017)**. We utilise the TreeSHAP approach detailed in the Xgboost package **(Lundberg, Erion, et al. 2018; T. Chen, T. He, et al. 2015)**.

8.2.13 Inferring networks

It has often been reported that associations between model parameters can carry additional information to the parameters **(Miao et al. 2011)**. In order to study this, we inferred associations between the parameters to build parameter-parameter interaction networks for different patient groups. We utilised a PCLRC-based approach tested for its robustness in **(Jahagirdar, Suarez-Diez, et al. 2019)**. The method is based on the iterative sampling of the association measures. For every iteration, 70% samples are randomly selected to infer the association matrix. The CLR algorithm **(Faith et al. 2007)** is then employed to determine the network edges with the help of a dynamic threshold that retains the top 30% of all edges. The entire procedure is repeated $K = 10^5$ times and a probabilistic measure is calculated for every edge. We apply a stringent threshold on the probability of edge likeliness (0.95) and reject every edge below this threshold. The edges in the network created have two values associated with them: (a) a measure of association and (b) a probability associated with the measure of association. The probability measure could be interpreted as the confidence level based on which one could approve or refuse the association between the two parameters.

We modified existing association measures to create an association measure to meet the following criteria.

- Capture the non-linear nature of the parameters and their interactions.
- Remove the influence of other parameters when considering the association between a parameter pair.
- Differentiate between positive and negative associations.

We define the association measure (a) between parameters X and Y given parameter

Table 8.3: The nomenclature used to describe the ratios and parameters in the chapter.

Ratio or parameter	Nomenclature
Ratio of Thrombomodulin & IL-4	Thrombomodulin $\Longleftrightarrow$ IL-4
Parameter γ when Thrombomodulin is the pro-inflammatory analyte & IL-4 the anti-inflammatory cytokine	γThrombomodulin $\Rightarrow$ IL-4
Parameter α_2 when Thrombomodulin is the pro-inflammatory analyte & IL-4 the anti-inflammatory cytokine	α_2Thrombomodulin $\Rightarrow$ IL-4

Z by modifying the existing PMI measure described in **(J. Zhao et al. 2016)** such that,

$$a(X;Y|Z) = \begin{cases} \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p^*(x|z)p^*(y|z)} & \text{if } \rho > 0 \\ -\sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p^*(x|z)p^*(y|z)} & \text{if } \rho < 0 \end{cases}. \tag{8.15}$$

where,

$$\rho = \frac{\text{cov}(X,Y)}{S_X \times S_Y}. \tag{8.16}$$

such that ρ represents the Pearson correlation measure, cov (X,Y) represents the covariance of the parameters X and Y, $p(x,y,z)$ is the joint probability distribution of parameters X, Y and Z. $p^*(x|z)$ and $p^*(y|z)$ was defined by **(J. Zhao et al. 2016)** as

$$p^*(x|z) = \sum_y p(x|z,y)p(y) \tag{8.17a}$$

$$p^*(y|z) = \sum_x p(y|z,x)p(x). \tag{8.17b}$$

8.2.14 Nomenclature

In order to improve the legibility of this manuscript, we have standardised the notation used for the relationship between the analytes being discussed. The notation is described in table 8.3. Note: The arrow is always pointing towards the anti-inflammatory analyte.

8.2.15 Software

Matlab **(MATLAB 2022)** and R **(Team et al. 2016)** were used for all the analysis performed in this manuscript. The ode45 in matlab solver was used to simulate the Baker model **(Shampine et al. 1997)**. The FME package in R was used for the sensitivity and identifiability analysis **(Soetaert et al. 2010)**. lhsdesign was used latin hypercube sampling and lsqnonlin **(Coleman et al. 1996)** was used for least-squares curve fitting. pROC package **(Robin et al. 2011)** was used for the ROC curves. The XGBosst R package **(T. Chen, T. He, et al. 2015)** was used for building the Xgboost models. Network Inference was coded in R modifying the PCLRC code available on https://www.systemsbiology.nl/download-page/ and PMI matlab code from **(J. Zhao**

et al. 2016). The supplementary can be found in
https://doi.org/10.5281/zenodo.8128358

8.3 Results

8.3.1 Parameters & their ability to classify patients

Studying the parameter space of the model, we found that α_2 & α_4 were the parameters that would influence the model the most as seen in the supplementary file S3 Chapter 8. When solving the mathematical equations in *Steady state assumption I* we found that we had the ability to pre-determine α_1 based on the multivariate sensitivities for the whole model. We also found that we could pre-determine γ when solving for α_4 based on their bi-variate sensitivity of 0.99 as seen in supplementary file S3 Chapter 8. On a similar assumption, we could pre-determine α_3 when solving for α_1 based on their bi-variate sensitivity of 0.71. A complete analysis of the parameter space can be found in the supplementary file S3 Chapter 8.

α_2 and α_4 values were calculated based on the first steady state assumption for each analyte pair per patient. ROC curves were constructed to assess the capabilities of these parameters to discriminate between NSTI and the various controls. A very high area under the curve (AUC) was taken as a measure of the parameter's ability to successfully distinguish NSTI from other controls. All the results are shown in the supplementary file S3 Chapter 8. The top 20 parameter pairs were selected for parameter estimation under steady state assumption II.

We used the distributions of accuracy and AUC values obtained from running 1000 Xgboost models to suss the information content possessed inside the pair-wise relationships of the analytes as compared to the analyte concentration alone. We use the accuracies and ROC-AUCs of the model performance in an ensemble to study the information value of the ratios and parameters. We calculate all model performance metrics on the testing datasets which were datasets not used in the training of the model itself. We report the best values, mean values and the 95% confidence intervals of both accuracy and AUC along with the top analyte/analyte pair in table 8.4. We also show the overview of the accuracies of the models in figure 8.4. In general, we find that as a trend the parameter α_4 shows the greatest ability to classify between patient outcomes, microbiological aetiology, and discriminating NSTI from the controls. The AUCs and accuracies are often higher than concentration values by a distinct margin. In most classifications, we find that ratios have higher performance metrics than concentrations alone with the exception of predicting the occurrence of septic shock. Even in the case of predicting septic shock α_4 parameter values have the highest performance metrics when we consider all the measures for pair-wise relationships studied and equivalent to the concentration metrics.

8.3.2 Discriminating NSTI from various controls

We build models discriminating NSTI from surgical controls, cellulitis and non-NSTI patients as described in *Patient cohorts and cytokine data*. The performance metrics for discriminating NSTI are some of the highest and the best models often achieved AUCs of 1 and accuracies of 100%. We found that MMP-8 and Pentraxin-3 had

Table 8.4

Models	Measure	Cytokine Concentrations	Ratios	Parameter α_2	Parameter α_4	All Parameters of the model
Descriminating NSTI						
NSTI vs Surgical Controls	Accuracy (%)	97.20[97.03-97.28]|(100)	98.19[98.12-98.33]|(100)	96.95[96.87-97.10]|(100)	97.48[97.28-97.50]|(100)	89.12[89.11-90.01]|(96.30)
	AUC	0.99[0.98-0.99]|(1)	0.95[0.94-0.95]|(1)	0.99[0.99-0.99]|(1)	0.96[0.95-0.96]|(1)	0.68[0.68-0.69]|(1)
	Top Feature	MMP-8	MMP-8 $\Longleftrightarrow$ Fas-Ligand	α_2CCL-5/RANTES $\Rightarrow$ MMP-8	α_4TNF$\alpha \Rightarrow$ Pentraxin-3	α_4G-CSF $\Rightarrow$ MPO
NSTI vs Cellulitis	Accuracy (%)	93.03[92.62-93.21]|(100)	94.54[94.35-94.69]|(100)	93.40[93.21-93.72]|(100)	95.08[94.93-95.27]|(100)	89.80[89.47-90.32]|(96.23)
	AUC	0.98[0.98-0.98]|(1)	0.99[0.99-0.99]|(1)	0.99[0.98-0.99]|(1)	0.99[0.98-0.99]|(1)	0.60[0.6-0.62]|(0.95)
	Top Feature	IL-6	C5/C5a $\Longleftrightarrow$ IL-1β	α_2I-α-1/COL1A1 $\Rightarrow$ IL-1β	α_4MMP-9 $\Rightarrow$ IL-1β	α_1Thrombomodulin $\Rightarrow$ CCL-4/MIP-1β
NSTI vs Non-NSTI	Accuracy (%)	88.47[88.20-88.81]|(98.15)	89.92[89.56-90.08]|(98.15)	89.67[89.32-89.86.43]|(98.15)	90.76[90.57 - 91.13]|(100)	88.48[88.72-89.21]|(94.44)
	AUC	0.94[0.93 - 0.94]|(1)	0.94[0.93 - 0.94]|(1)	0.94[0.94-0.94]|(1)	0.96[0.96 -0.96]|(1)	0.62[0.65-0.66]|(0.99)
	Top Feature	Thrombomodulin	CCL-4/MIP-1$\beta \Longleftrightarrow$ Thrombomodulin	α_2IL-4 $\Rightarrow$ Thrombomodulin	α_4IL-4 $\Rightarrow$ Thrombomodulin	α_4IL-6 $\Rightarrow$ ICAM-1
Patient Outcome						
Mortality	Accuracy (%)	81.22[80.80-81.96]|(98)	81.89[81.80-82.59]|(96)	81.73[80.88-81.79]|(94)	81.43[81.26-81.96]|(96)	80.21[80.19-82.19]|(90)
	AUC	0.79[0.78-0.79]|(0.99)	0.83[0.82-0.83]|(1)	0.76[0.75-0.76]|(0.99)	0.81[0.81-0.82]|(1)	0.66[0.66-0.67]|(0.92)
	Top Feature	MMP-1	C5/C3a $\Longleftrightarrow$ MMP-1	α_2CXCL-10/IP-10 $\Rightarrow$ Collagen-IVα1	α_4MMP-1 $\Rightarrow$ C5/C5a	α_1Thrombomodulin $\Rightarrow$ S100A8
Septicshock occurrence	Accuracy (%)	72.19[71.90-72.63]|(90)	72[71.54-72.26]|(88)	70.79[70.29-71.02]|(88)	72.26[71.77-72.47]|(88)	50.01[49.39-53.89]|(70)
	AUC	0.80[0.80-0.80]|(0.96)	0.80[0.79-0.80]|(0.96)	0.78[0.78-0.79]|(0.95)	0.80[0.79-0.80]|(0.96)	0.53[0.53-0.53]|(0.71)
	Top Feature	S100A8	C5/C5a $\Longleftrightarrow$ IL-6	α_2Resistin $\Rightarrow$ IL-6	α_4C5/C5a $\Rightarrow$ IL-6	α_2Pentraxin-3 $\Rightarrow$ CCL-5/RANTES
Microbial Aetiology						
Mono vs Poly	Accuracy (%)	76.47[76.37-77.05]|(96)	80[79.51-80.17]|(96)	75.58[74.96-75.64]|(92)	78[77.70-78.39]|(94)	53.78[53.18-53.94]|(74)
	AUC	0.86[0.86-0.87]|(0.99)	0.88[0.87-0.88]|(0.99)	0.85[0.84-0.85]|(0.98)	0.86[0.85-0.86]|(0.98)	0.56[0.55-0.56]|(0.76)
	Top Feature	CXCL-10/IP-10	CXCL-10/IP-10 $\Longleftrightarrow$ IL-4	α_2C5/C5a $\Rightarrow$ CXCL-10/IP-10	α_4CXCL-10/IP-10 $\Rightarrow$ CCL-4/MIP-1β	α_3IL-6 $\Rightarrow$ IL-13
GAS vs SD	Accuracy (%)	74.3[73.61-75.0]|(95.83)	79.17[78.54-79.51]|(95.83)	77.8[77.08-78.46]|(95.83)	78[77.47-78.53]|(95.83)	77.02[76.27-77.77]|(95.83)
	AUC	0.90[0.89-0.90]|(1)	0.91[0.90-0.91]|(1)	0.83[0.82-0.84]|(1)	0.93[0.92-0.93]|(1)	0.67[0.67-0.68]|(0.93)
	Top Feature	CXCL-10/IP-10	CXCL-10/IP-10 $\Longleftrightarrow$ IL-6	α_2IL-33 $\Rightarrow$ CXCL-10/IP-10	α_4CXCL-10/IP-10 $\Rightarrow$ IL-2	γThrombomodulin $\Rightarrow$ MMP-8

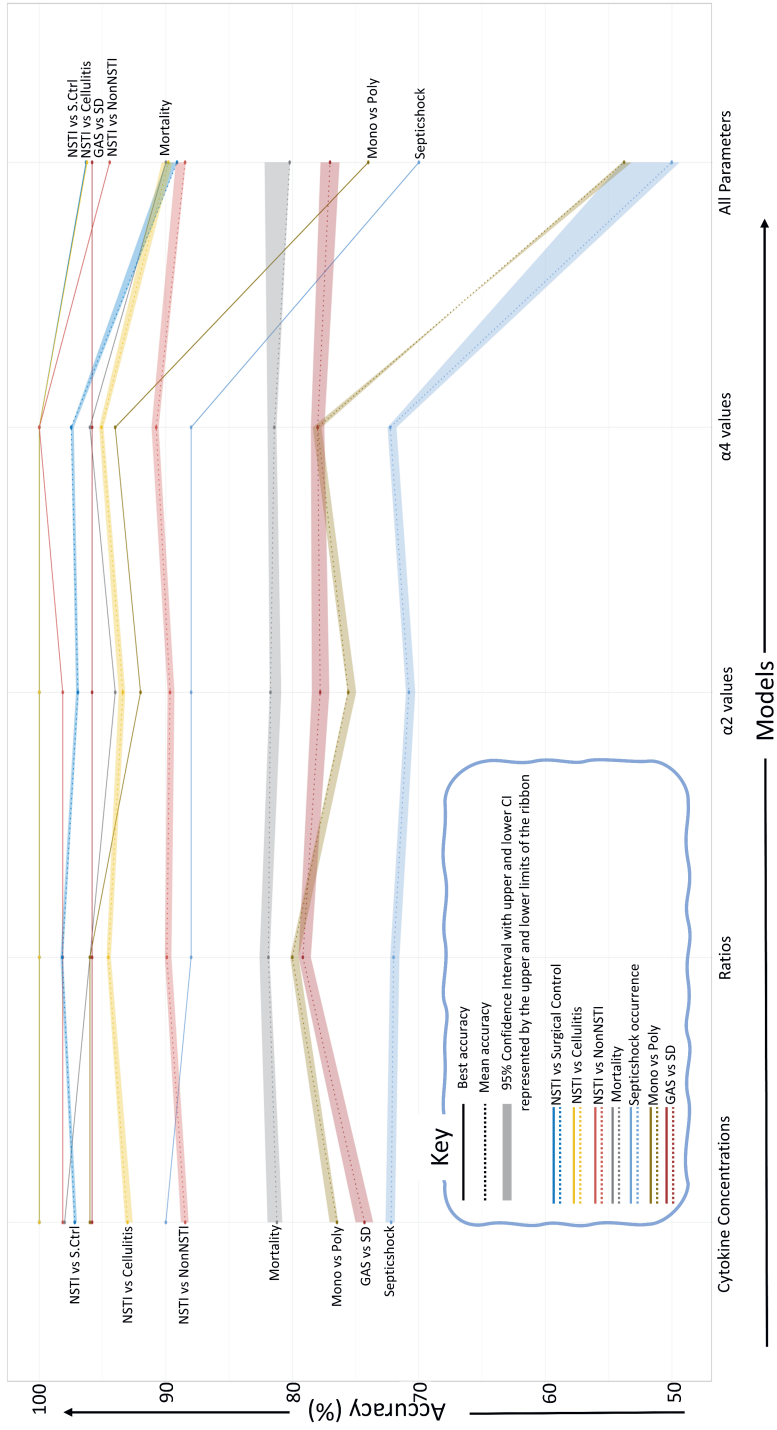

Accuracy (%)
100
90
80
70
60
50
Models
Cytokine Concentrations
Ratios
α2 values
α4 values
All Parameters
NSTI vs S.Ctrl
NSTI vs Cellulitis
NSTI vs NonNSTI
Mortality
Septicshock
Mono vs Poly
GAS vs SD
Key
Best accuracy
Mean accuracy
95% Confidence Interval with upper and lower CI represented by the upper and lower limits of the ribbon
NSTI vs Surgical Control
NSTI vs Cellulitis
NSTI vs NonNSTI
Mortality
Septicshock occurrence
Mono vs Poly
GAS vs SD

Figure 8.5

Table 8.4: Overview of the results from running the Xgboost models. Each result contains the properties of the distribution created by running 1000 models. Each cell in the table gives the mean value followed by the 95% confidence intervals (CIs) in square brackets and the best value in parenthesis. As an example, we report the accuracy values 97.20[97.03-97.28](100) when running the models for classification between NSTI and surgical controls using the cytokine concentration values. This is reporting the cumulative result of 1000 models where the accuracy of the best-performing model is 100%, the mean of all accuracies is 97.20%, the lower CI is 97.03% and the higher CI is 97.28%. The Area under the curves (AUC) or the ROC curves depicting the model performances are reported in the same format.

Figure 8.5: The overall results of all the Xgboost models in terms of the accuracies. The best, mean and the 95% confidence intervals of the distribution of 1000 accuracies representing performance of 1000 Xgboost models are shown by bold, dotted lines and ribbons. The X-axis represents the data on which the models are built and the Y-axis represents the accuracy in percentage.

the most effect when discriminating between NSTI and surgical controls using concentration values with IL-6 and Collagen-IVα1 also having a smaller effect. When we study the pair-wise relationships, we find that MMP-8 $\iff$ Fas-Ligand, IL-6 $\iff$ IL-4, α_2CCL-5/RANTES $\Rightarrow$ MMP-8, α_4TNFα $\Rightarrow$ Pentraxin-3, α_4G-CSF $\Rightarrow$ MPO, α_1Pentraxin-3 $\Rightarrow$ CCL-5/RANTES & α_3G-CSF $\Rightarrow$ CXCL-10/IP-10 were the parameters with significant importance on the models. The SHAP plots of some of the selected concentrations and parameters are shown in figure 8.6. All the SHAP plots can be found in supplementary file S3 Chapter 8.

When classifying between NSTI and cellulitis patients, we found that the concentrations of IL-6, IL-1β, CXCL-8/IL-8 & IL-23 had the most effect with smaller effects by the concentrations of IL-33 & Thrombomodulin. When we use the measures for pair-wise relationships, we find C5/C5a $\iff$ IL-1β, α_2C5/C5a $\Rightarrow$ CCL-4/MIP-1β, α_2I-α-1/COL1A1 $\Rightarrow$ IL-1β, α_4MMP-9 $\Rightarrow$ IL-1β, α_1Thrombomodulin $\Rightarrow$ CCL-4/MIP-1β, α_3Thrombomodulin $\Rightarrow$ S100A8 & α_1G-CSF $\Rightarrow$ MPO having effect on the outcome of the models. The effect of the concentration, ratios and parameters on the SHAP values are shown in figure 8.7.

Classifying Non-NSTI and NSTI patients had a relatively lower accuracy compared to the other controls, however, the performance metrics were still extremely good. We find that the concentration of Thrombomodulin (Gain = 0.674) had a significantly higher effect than any other analyte (Fas-Ligand; gain = 0.054). When we look at the SHAP plot, the model predicts that patients with Thrombomodulin concentration <~8000 pg/ml are Non-NSTI patients and patients with >~8000pg/ml are NSTI patients. When we study the pair-wise relationships, they are also dominated by associations with Thrombomodulin. We find that CCL-4/MIP-1β $\iff$ Thrombomodulin, TNFα $\iff$ Fas-Ligand, α_2IL-4 $\Rightarrow$ Thrombomodulin, α_2Fas-Ligand $\Rightarrow$ Thrombomodulin, α_2S100A8 $\Rightarrow$ Thrombomodulin, α_4IL-4 $\Rightarrow$ Thrombomodulin, α_4Fas-Ligand $\Rightarrow$ TNFα, γTNFα $\Rightarrow$ CXCL-8/IL-8, α_2G-CSF $\Rightarrow$ CXCL-10/IP-10, α_2Thrombomodulin $\Rightarrow$ CCL-4/MIP-1β & γG-CSF $\Rightarrow$ MPO have significant effect on the classifications. When we study the SHAP plots in figure 8.8, we see similar switches between the model predicting NSTI and NonNSTI patients. The model pre-

Table 8.5: The concentrations, ratios and parameters of importance isolated from the feature importance of the Xgboost models built on discriminating NSTI, microbial aetiology, and patient outcome.

Parameters	Discriminating NSTI		Microbial Aetiology		Patient Outcome	
	NSTI vs Surgical controls	NSTI vs Cellulitis	NSTI vs Non-NSTI	Mono vs Poly	GAS vs SD	Mortality	Septic shock
Concentrations	MMP-8	IL-6	Thrombomodulin	CXCL-10/IP-10	CXCL-10/IP-10	MMP-1	S100A8
		IL-1β		MMP-9	E-Selectin	E-Selectin	C5/C5a
		CXCL-8/IL-8			IL-33		
Ratios	IL-6 $\Longleftrightarrow$ IL-4	C5/C5a $\Longleftrightarrow$ IL-1β	CCL-4/MIP-1β $\Longleftrightarrow$ Thrombomodulin	Fas-Ligand $\Longleftrightarrow$ VCAM-1	CXCL-10/IP-10 $\Longleftrightarrow$ IL-6	C5/C5a $\Longleftrightarrow$ MMP-1	CCL-4/MIP-1β $\Longleftrightarrow$ IL-4
			TNFα $\Longleftrightarrow$ Fas-Ligand	CCL-4/MIP-1β $\Longleftrightarrow$ CXCL-10/IP-10	C5/C5a $\Longleftrightarrow$ CXCL-10/IP-10	S100A9 $\Longleftrightarrow$ CXCL-8/IL-8	C5/C5a $\Longleftrightarrow$ IL-6
				CXCL-10/IP-10 $\Longleftrightarrow$ IL-4		Thrombomodulin $\Longleftrightarrow$ Galectin-3	
Parameters	α_2CCL-5/RANTES $\Rightarrow$ MMP-8	α_2C5/C5a $\Rightarrow$ CCL-4/MIP-1β	α_2IL-4 $\Rightarrow$ Thrombomodulin	α_4C5/C5a $\Rightarrow$ CXCL-10/IP-10	α_2IL-33 $\Rightarrow$ CXCL-10/IP-10	α_2MMP-1 $\Rightarrow$ E-Selectin	α_2Resistin $\Rightarrow$ IL-6
	α_4TNFα $\Rightarrow$ Pentraxin-3	α_4IL-1β $\Rightarrow$ MMP-9	α_2Fas-Ligand $\Rightarrow$ Thrombomodulin	α_4CXCL-10/IP-10 $\Rightarrow$ CCL-4/MIP-1β	α_2IL-23 $\Rightarrow$ IL-22	α_2CXCL-10/IP-10 $\Rightarrow$ Collagen-IVa1	α_4IL-13 $\Rightarrow$ S100A8
	α_3G-CSF $\Rightarrow$ CXCL-10/IP-10	α_1Thrombomodulin $\Rightarrow$ CCL-4/MIP-1β	α_2S100A8 $\Rightarrow$ Thrombomodulin	α_3IL-6 $\Rightarrow$ IL-13	α_4CXCL-10/IP-10 $\Rightarrow$ IL-2	α_4E-Selectin $\Rightarrow$ Collagen-IVa1	α_4Thrombomodulin $\Rightarrow$ S100A8
	α_1Pentraxin-3 $\Rightarrow$ CCL-5/RANTES	α_3Thrombomodulin $\Rightarrow$ S100A8	α_4IL-4 $\Rightarrow$ Thrombomodulin	α_3IL-1β $\Rightarrow$ CCL-5/RANTES	γThrombomodulin $\Rightarrow$ MMP-8	α_1Thrombomodulin $\Rightarrow$ S100A8	α_2Pentraxin-3 $\Rightarrow$ CCL-5/RANTES
		α_1G-CSF $\Rightarrow$ MPO	α_4Fas-Ligand $\Rightarrow$ TNFα	α_1G-CSF $\Rightarrow$ CXCL-10/IP-10	α_1TNFα $\Rightarrow$ Fas-Ligand	γTNFα $\Rightarrow$ Fas-Ligand	
		α_2Pentraxin-3 $\Rightarrow$ CCL-5/RANTES	γTNFα $\Rightarrow$ CXCL-8/IL-8		α_3G-CSF $\Rightarrow$ CXCL-10/IP-10		
		α_4TNFα $\Rightarrow$ Fas-Ligand	α_2G-CSF $\Rightarrow$ CXCL-10/IP-10				
			α_2Thrombomodulin $\Rightarrow$ CCL-4/MIP-1β				
			γG-CSF $\Rightarrow$ MPO				

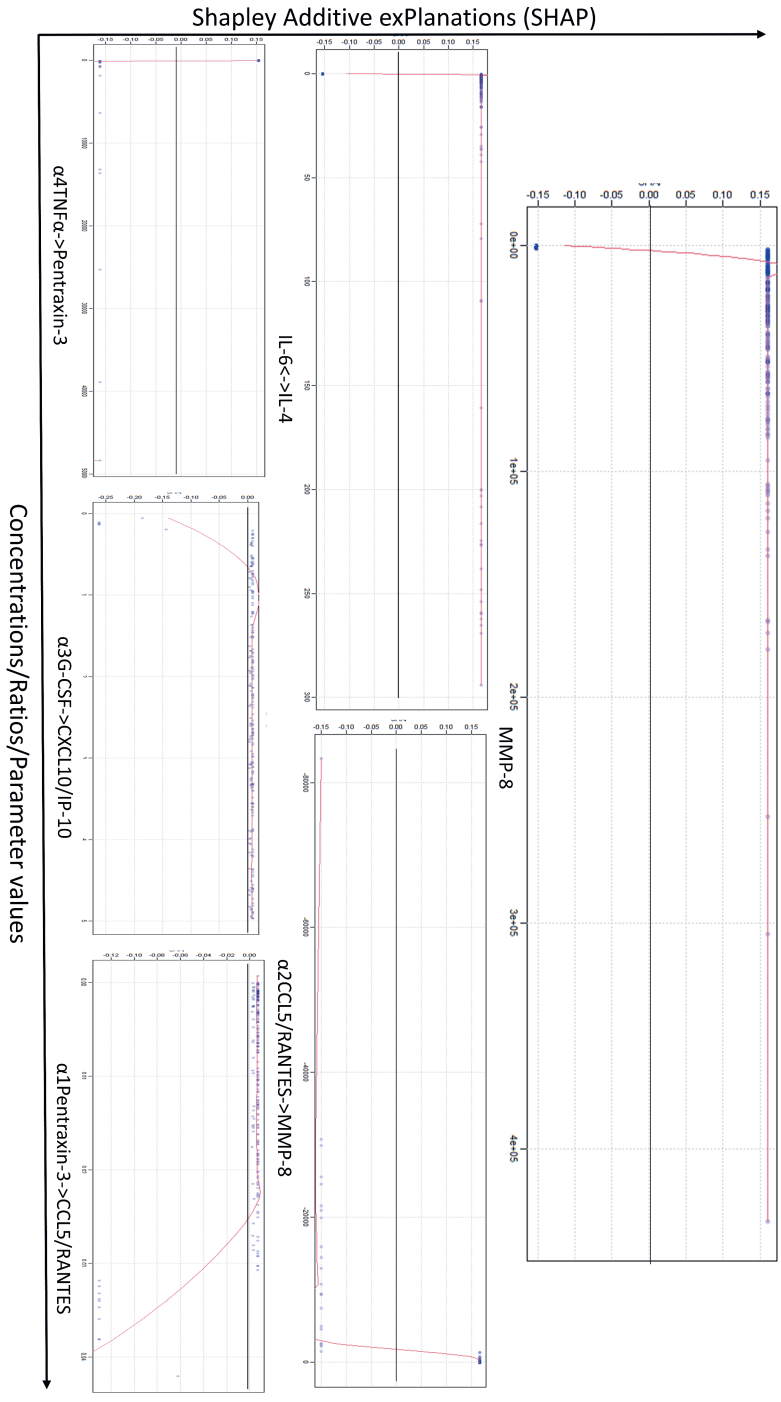

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
α4TNFα->Pentraxin-3
IL-6<->IL-4
MMP-8
α3G-CSF->CXCL10/IP-10
α2CCL5/RANTES->MMP-8
α1Pentraxin-3->CCL5/RANTES

Figure 8.6

Figure 8.6: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between NSTI and surgical controls. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as NSTI patient or surgical control. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient as NSTI. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient as surgical control. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

diction switches for CCL-4/MIP-1β $\Longleftrightarrow$ Thrombomodulin at a ratio of $\sim$0.10 and for α_2IL-4 $\Rightarrow$ Thrombomodulin at the parameter value of $\sim$12.

8.3.3 Classifying patients based on patient outcome

We chose to differentiate NSTI patients by mortality and septic shock occurrence. In this case, we only included the patients that were diagnosed with NSTI. In general, we find the prediction accuracies to be lower than when discriminating NSTI from controls. When classifying by mortality, we find that MMP-1, E-Selectin, C5/C5a, Galectin-3, CXCL-8/IL-8 & Collagen-IVα1 had a strong effect on the classification. Select SHAP plots are shown in figure 8.9.

When classifying based on whether or not the patient suffered a septic shock, we find that the concentration of S100A8 had a significant effect on the classification with the concentration C5/C5a, Resistin & MMP-8 also having an effect albeit lower than S100A8. The model splits patients with concentration >$\sim$900 pg/ml to have had septic shock as seen in the figure 8.9. When studying the pair-wise relationships, we find that CCL-4/MIP-1β $\Longleftrightarrow$ IL-4, C5/C5a $\Longleftrightarrow$ IL-6, α_2Resistin $\Rightarrow$ IL-6, α_4IL-13 $\Rightarrow$ S100A8, α_4Thrombomodulin $\Rightarrow$ S100A8 & α_2Pentraxin-3 $\Rightarrow$ CCL-5/RANTES had significant effects on the classification.

8.3.4 Classifying patients based on microbial aetiology

We focus on the following two distinctions that clinicians have been interested in. Firstly, using the analyte concentrations, ratios and parameter values to differentiate between mono-microial infections and poly-microbial infections and second, to differentiate between patients whose infections were caused by Group A Streptococcus (GAS) and caused by *S. dysgalactiae* (GCS, GGS, SD). When we used the model to classify between mono- and poly-microbial infections, we found that concentration of CXCL-10/IP-10 had the most significant effect. Concentrations of MMP-9, C5/C5a, Fas-Ligand also had effects on the classifications. When we study the measures of relationships between the analytes, we find that CCL-4/MIP-1β $\Longleftrightarrow$ CXCL-10/IP-10, CXCL-10/IP-10 $\Longleftrightarrow$ IL-4, Fas-Ligand $\Longleftrightarrow$ VCAM-1, α_2C5/C5a $\Rightarrow$ CXCL-10/IP-10, α_4CXCL-10/IP-10 $\Rightarrow$ CCL-4/MIP-1β, α_3IL-6 $\Rightarrow$ IL-13, α_3IL-1β $\Rightarrow$ CCL-5/RANTES, α_1G-CSF $\Rightarrow$ CXCL-10/IP-10 to have a significant effect on the classification. The model particularly shows a difference with CCL-4/MIP-1β $\Longleftrightarrow$ CXCL-10/IP-10. The model classifies ratios <$\sim$4 as mono-microbial infections as seen in figure 8.11.

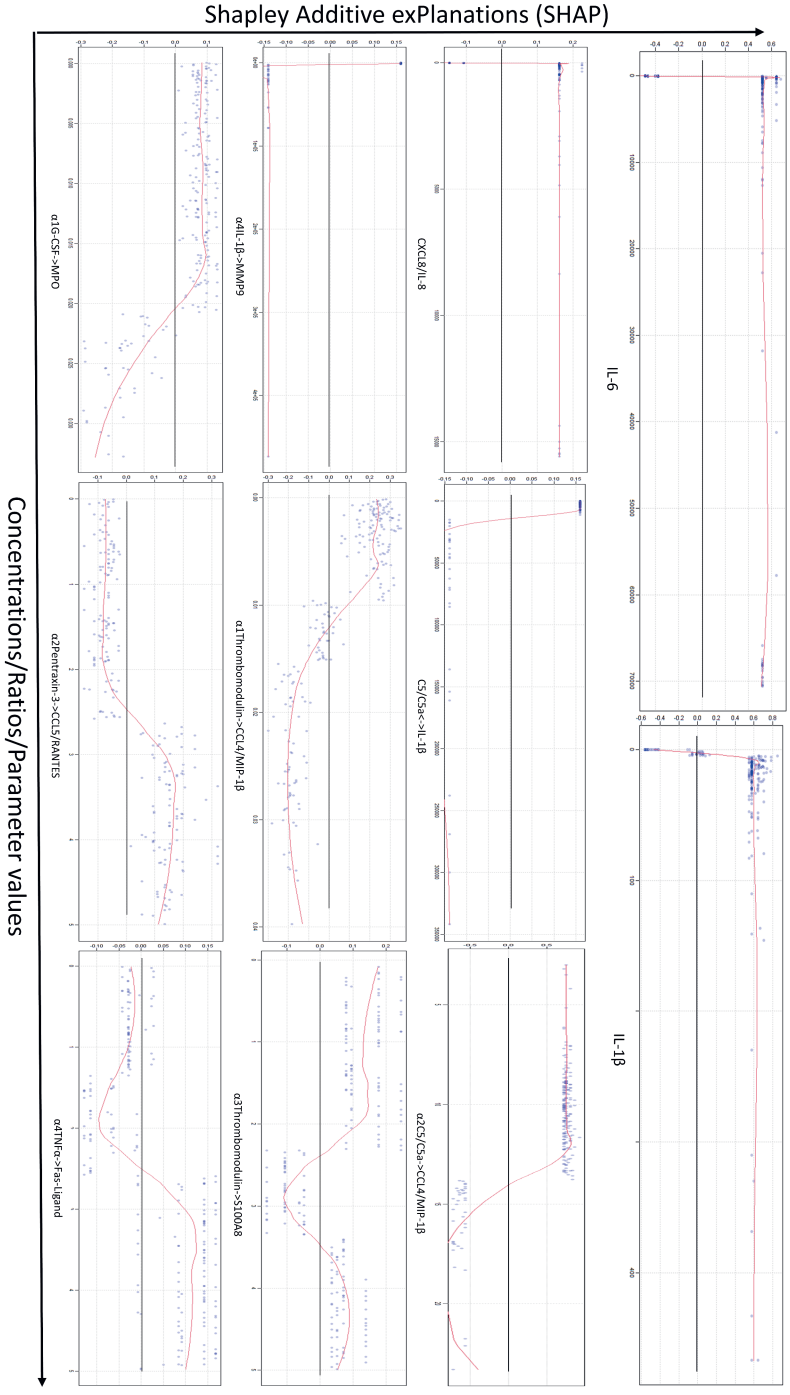

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
α1G-CSF->MPO
α4IL-1β->MMP9
CXCL8/IL-8
IL-6
α2Pentraxin-3->CCL5/RANTES
α1Thrombomodulin->CCL4/MIP-1β
C5/C5a<->IL-1β
α4TNFα->Fas-Ligand
α3Thrombomodulin->S100A8
α2C5/C5a->CCL4/MIP-1β
IL-1β

Figure 8.7

Figure 8.7: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between NSTI and surgical controls. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as NSTI patient or cellulitis patient. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient as NSTI. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient as having cellulitis. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

When we classify based on the strain that was found to be the cause of the infection, we find that once again the concentration of CXCL-10/IP-10 having a strong effect. However, in this instance we find the concentrations of E-Selectin and IL-33 also having a strong effect followed by IL-17A and IL-23 having a less accentuated but significant effect on the classification. When we study the relationship measures, we find CXCL-10/IP-10 $\iff$ IL-6, C5/C5a $\iff$ CXCL-10/IP-10, α_4CXCl-10/IP-10 $\Rightarrow$ IL-2, α_2IL-33 $\Rightarrow$ CXCL-10/IP-10, α_2IL-23 $\Rightarrow$ IL-22, γThrombomodulin $\Rightarrow$ MMP-8, α_1TNFα $\Rightarrow$ Fas-Ligand & α_3G-CSF $\Rightarrow$ CXCL-10/IP-10 to have significant effects on the classification. Graphs on selected features are shown in figure 8.11.

8.3.5 Associations between parameters

We study the associations between the estimated parameters from the patients and controls and the networks representing (A) NSTI patients, (B) non-NSTI patients, (C) Cellulitis patients and (D) surgical controls are visualised in figure 8.13. We see significantly higher activity from the higher number of associations in NSTI patients compared to the controls. Non-NSTI and cellulitis patients also display a slightly higher activity than surgical controls. We observe a positive association between α_2Thrombomodulin $\Rightarrow$ S100A8 and α_4IL-1β $\Rightarrow$ IL-1α in cellulitis patients that is contrasted with a negative association of α_3Thrombomodulin $\Rightarrow$ S100A8 and α_4IL-1β $\Rightarrow$ IL-1α in NSTI patients. NSTI patients show a further flurry of negative associations with α_3Thrombomodulin $\Rightarrow$ MMP-8 and α_4Thrombomodulin $\Rightarrow$ IL-6 being negatively associated with α_3IL-1β $\Rightarrow$ IL-12p70 and γIL-6 $\Rightarrow$ IL-13 respectively. α_4IL-1β $\Rightarrow$ IL-1α is also negatively associated with α_3S100A8 $\Rightarrow$ CXCL-10/IP-10. No such associations are observed in the non-NSTI network with respect to analyte pairs involving Thrombomodulin. Instead, we see a negetive association between α_4TNFα $\Rightarrow$ Fas-Ligand and α_4Pentraxin-3 $\Rightarrow$ CCL-5/RANTES and a positive association between γG-CSF $\Rightarrow$ MPO and α_2IL-1β $\Rightarrow$ IL-12p70. Along with the properties of the data, the characteristics introduced by the baker model itself can also be seen in the positive associations between α_3 and γ parameters of the same analyte pair.

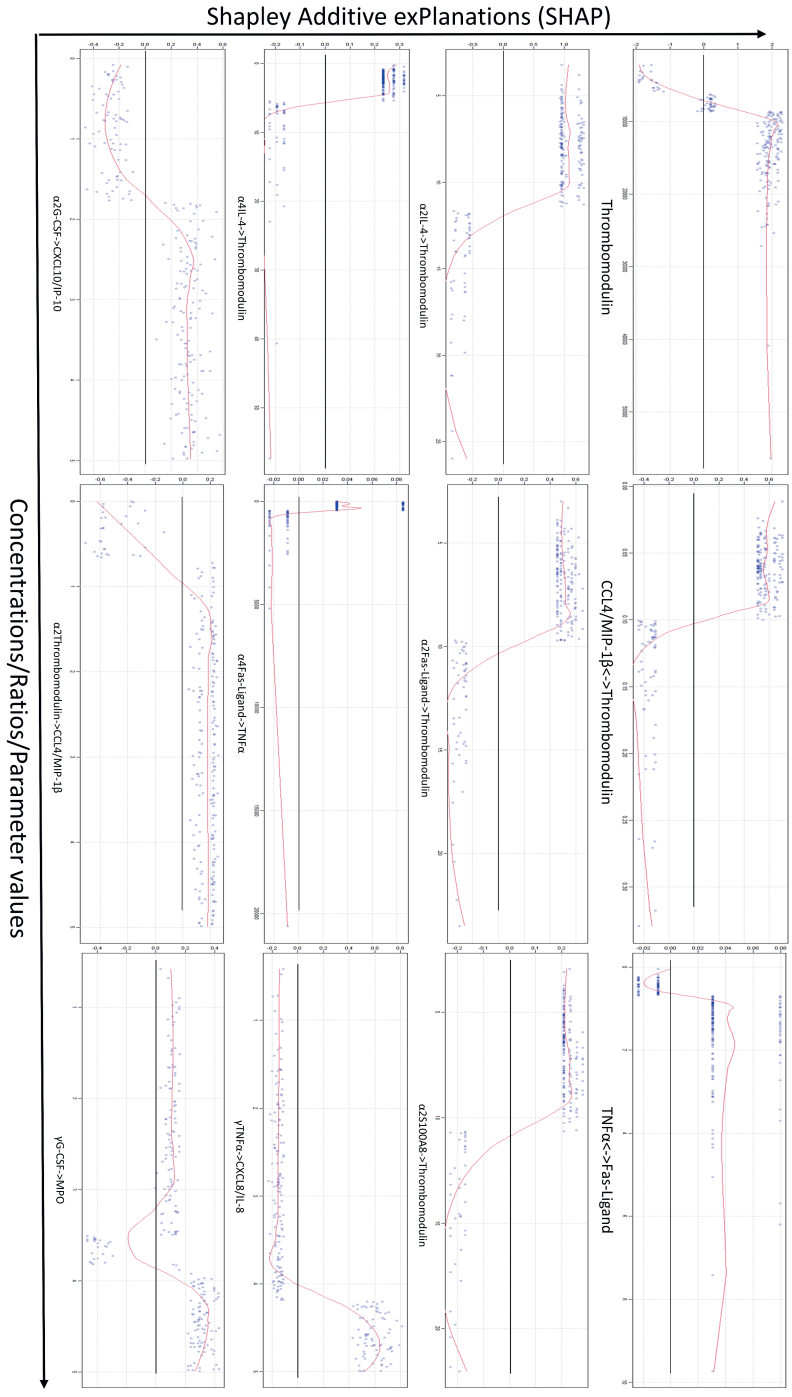

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
α2G-CSF->CXCL10/IP-10
α4IL-4->Thrombomodulin
α2IL-4->Thrombomodulin
Thrombomodulin
α2Thrombomodulin->CCL4/MIP-1β
α4Fas-Ligand->TNFα
α2Fas-Ligand->Thrombomodulin
CCL4/MIP-1β<->Thrombomodulin
γG-CSF->MPO
γTNFα->CXCL8/IL-8
α2S100A8->Thrombomodulin
TNFα<->Fas-Ligand

Figure 8.8

Figure 8.8: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between NSTI and non-NSTI. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as NSTI patient or non-NSTI patient. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient as NSTI. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient as non-NSTI. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

8.4 Discussion

8.4.1 Concentrations, ratios or model parameters?

We explore the distribution characteristics of the performance measures obtained from running a 1000 XGbosst models to assess the information content in the ratios and model parameters and compare it with the original concentration values measured in the plasma of patients on the day of admission to the hospital. The mean values, 95% confidence intervals and the best values of the performance measures are reported in table 8.4. In most of the two-group classifications we perform, we find that both the ratios and model parameters have higher accuracies and AUCs than the concentration values with the exception of septic shock occurrence. Even when classifying based on the occurrence of septic shock, the accuracy of classification based on α_4 values is similar to the accuracy of classification based on concentration values. Overall, accuracies were highest when classifications were performed using α_4 parameter values. The classification accuracies were slightly lower when we used all the estimated parameters together when discriminating NSTI cases and more significantly lower when classifying based on septic shock occurrence. The drop in accuracies in this scenario is partially expected on account of (a) increased introduction of noise due to the model characteristics and the estimation procedure, and (b) due to the pre-selection of 20 analyte pairs leading to information depletion. The pre-selection was necessary to make the exploratory strategy computationally viable. The overall small increase in classification accuracy in ratios and parameters is not the only informational gain that has occured from using such a strategy, but also, the increase in information regarding the pair-wise relationships between the analytes provide valuable clues relating to the underlying mechanisms. The information provided by these pair-wise relationship can also be valuable when stratifying patients with similar responses.

8.4.2 Discriminating NSTI

MMP-8 (Matrix Metalloproteinase-8), also known as collagenase-2, is an proteolytic enzyme that belongs to a family of zinc-dependant endopeptidases. It is primarily involved in the breakdown of extracellular matrix components, such as collagen I, II, III. fibronectin, aggrecan and ovostatin. It is produced and secreted by a variety of cells, including neutrophils, activated macrophages, chondrocytes fibroblasts, ep-

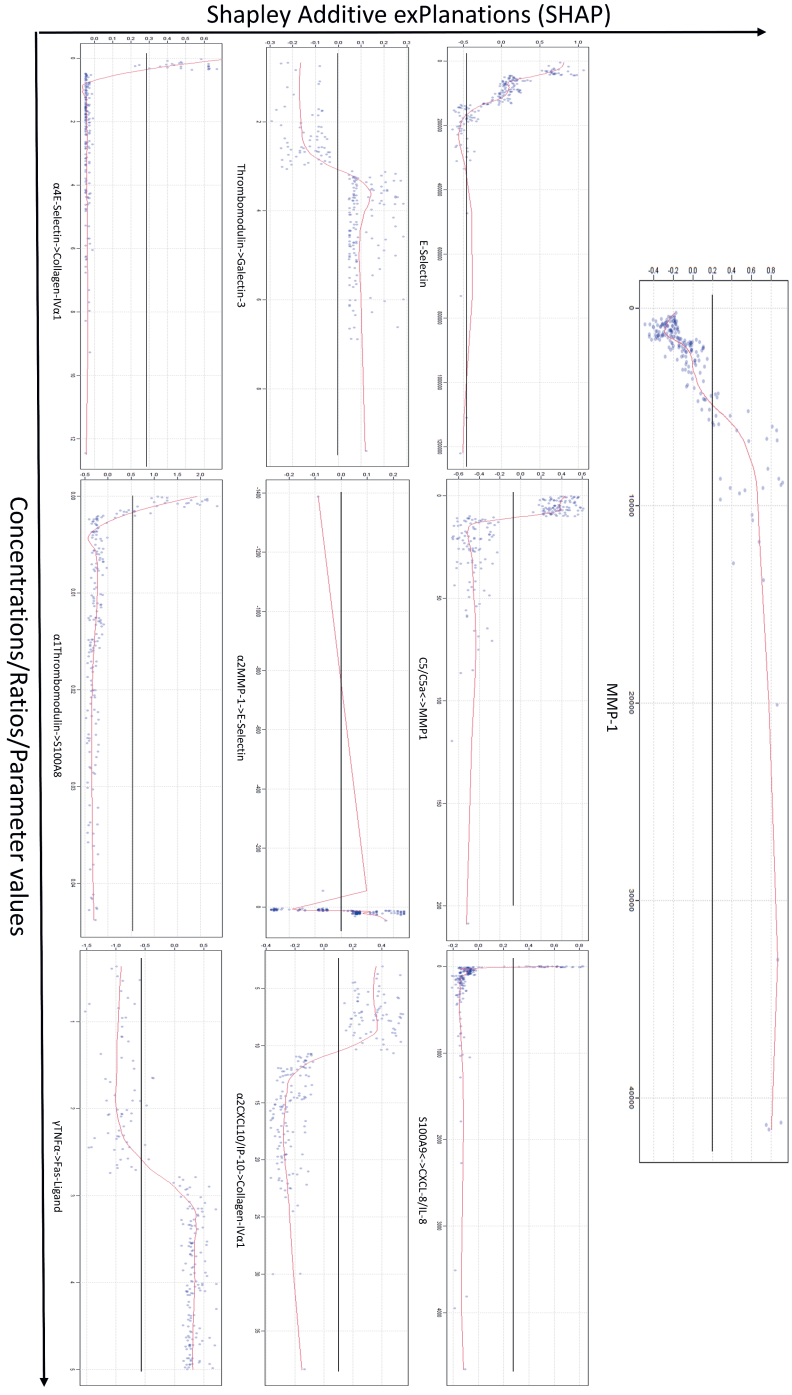

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
α4E-Selectin->Collagen-IVα1
Thrombomodulin->Galectin-3
E-Selectin
MMP-1
α1Thrombomodulin->S100A8
α2MMP-1->E-Selectin
C5/C5a<->MMP1
γTNFα->Fas-Ligand
α2CXCL10/IP-10->Collagen-IVα1
S100A9<->CXCL-8/IL-8

Figure 8.9

Figure 8.9: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between patients based on mortality. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as non-survivor patient or survival patient. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient as non-survivor. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient as survivor. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

ithelial and endothelial cells and smooth muscle cells **(Laronha et al. 2020; Luchian et al. 2022)**. The higher concentration of MMP-8 in NSTI compared to the surgical controls suggests a greater propensity towards wound healing and tissue remodelling **(Laronha et al. 2020; Luchian et al. 2022)**. Other analytes that had a significant effect in descriminating between the surgical controls and NSTI are Pentraxin-3 and IL-6. Pentraxin-3, as a soluble pattern recognition receptor also promotes phagocytosis by macrophages and neutrophils **(Koussih et al. 2021)**. IL-6 is also secreted by macrophages in response pathogen-associated molecular patterns (PAMPs) and is responsible for stimulating the production on neutrophils among it's many other functions **(Mihara et al. 2012)**. The top differentiators all seem to hint towards a greater presence of neutrophils and macrophages in NSTI patients compared to surgical controls. Not only has the presence of neutrophils and macrophages at the site of infection in NSTI been reported but also their contribution towards tissue inflammation and damage through processes like neutrophil degranulation has been studied **(Siemens, Snäll, et al. 2020)**.

Fas-Ligand is a surface protein on immune cells that induces apoptosis **(Nagata 1999)**. There is no obvious interaction between MMP-8 and Fas-Ligand, yet we find that the ratio of the two is a very strong differentiator between NSTI and surgical controls. Korpi et al. demonstrate while studying wound healing in tooth-extracted mice that Fas-Ligand can be a substrate for MMP-8 and indicate towards spatial, temporal production and processing of Fas-Ligand by MMP-8 **(Korpi et al. 2009)**. Even though MMP-8 $\iff$ Fas-Ligand is a strong indicator, it does not necessarily point towards the existence of a direct interaction between two. Similarly, IL-6 $\iff$ IL-4 seem to be a strong differentiator, however IL-6 and IL-4 play a complete opposite role, where IL-6 is a pro-inflammatory cytokine which is produced in response to tissue damage and IL-4 is an anti-inflammatory cytokine produced mainly by T helper 2 (Th2) cells.

Concentrations of IL-6, IL-1β, CXCL-8/IL-8 & IL-23 had an effect when classifying between NSTI and cellulitis patients. All of these analytes were also found to have a very high AUC and mean decrease in gini (MDG) in the ROC analysis and Random Forest models run while differentiating between streptococcal NSTI and cellulitis **(Rath et al. 2023)**. Discriminatory abilities of C5/C5a $\iff$ IL-1β and C5/C5a $\iff$ IL-6 are within reason as C5/C5a is formed during the activation of the complement cascade and the regulatory effects of IL-1β and IL-6 on C5a in other inflammatory situations has been shown **(Khameneh et al. 2017; Ignatius et al. 2011)**. Thrombomodulin also appears to have some ability to differentiate between cellulitis and

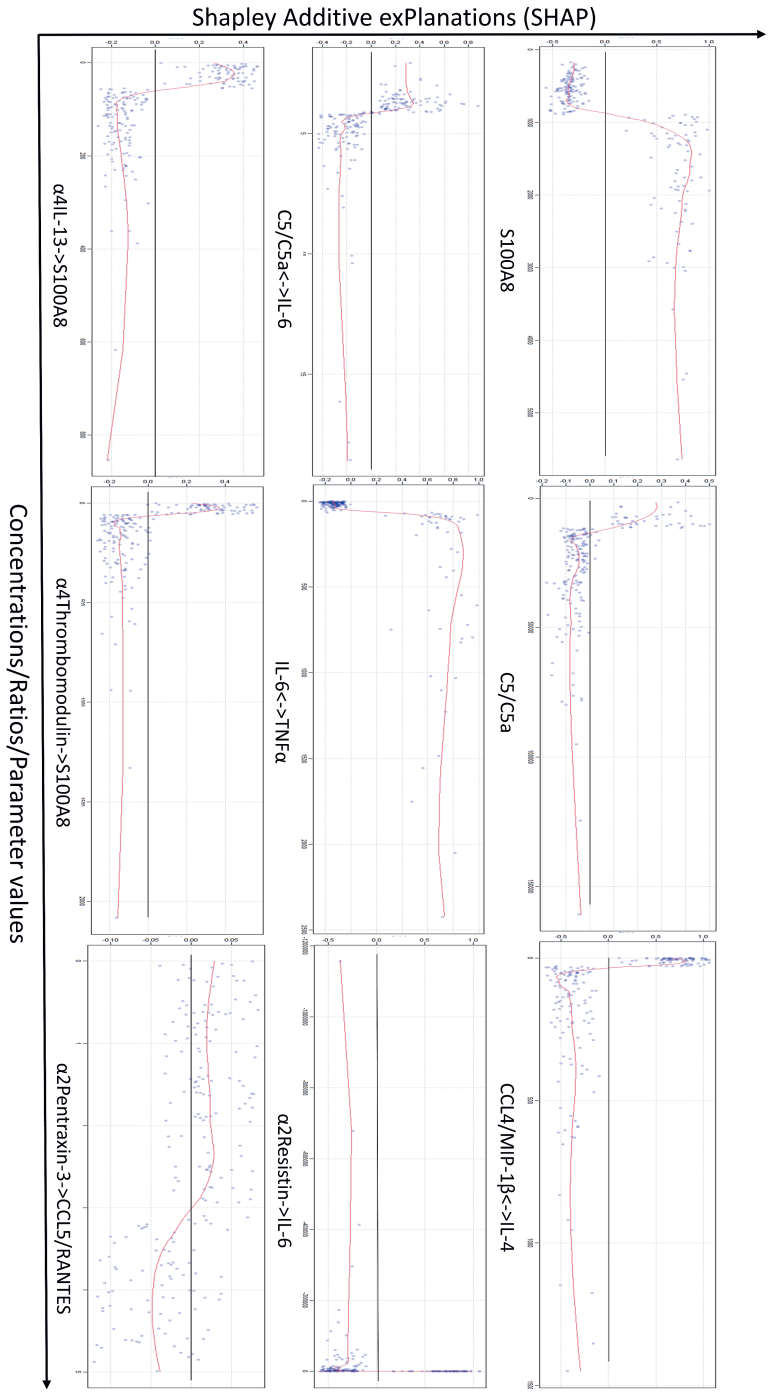

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
α4IL-13->S100A8
C5/C5a<->IL-6
S100A8
α4Thrombomodulin->S100A8
IL-6<->TNFα
C5/C5a
α2Pentraxin-3->CCL5/RANTES
α2Resistin->IL-6
CCL4/MIP-1β<->IL-4

Figure 8.10

Figure 8.10: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between patients based on septic shock occurrence. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as patient with septic shock occurrence or patient with septic shock non-occurrence. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient with septic shock occurrence. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient with the non-occurrence of septic shock. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

NSTI patients although not as pronounced as when differentiating between NSTI and Non-NSTI patients. Rath et al also come to a similar conclusion where they find Thrombomodulin's significance to be just below the p-value threshold in their feature importance when comparing cellulitis and streptococcal NSTI patients **(Rath et al. 2023)**. However, it is plausible that the lack of more severe cellulitis cases in the data may have an effect on this.

Thrombomodulin is a transmembrane multidomain glycoprotein receptor that plays a key role in regulating blood co-agulation and inflammation. It binds thrombin and promotes the activation of protein C that enables anti-inflammatory effects **(Weiler et al. 2003)**. Thrombomodulin alone has a significant discriminatory ability between NSTI and Non-NSTI in the Xgboost model. Thrombomodulin was also reported as a discriminatory plasma biomarker for NSTI recently **(Medina et al. 2021)**. Thrombomodulin has also been reported to play a critical role in co-agulation and inflammatory pathways crucial in sepsis patients **(Levi and Van Der Poll 2012)**. There has been experimental evidence showing significant downregulation of Thrombomodulin in sepsis patients caused by pro-inflammatory cytokines TNFα and IL-1 accompanied by diminished protein C activation **(Levi and Poll 2008)** and Yamakawa et al reported a moderate trend in the reduction of mortality of sepsis-induced DIC patients with recombinant human soluble Thrombomodulin **(Yamakawa, Aihara, et al. 2015)**. This could potentially explain the lower concentration of Thrombomodulin seen in Non-NSTI patients compared to the NSTI patients. LP Medina et al. report the discriminatory threshold for Thrombomodulin at 7566.85 pg/ml in their ROC analysis for NSTI and Non-NSTI patients (similar to what we see in figure 8.7 using the Xgboost model) and also found Thrombomodulin as a significant discriminator between NSTI and sepsis patients in their validation cohort **(Medina et al. 2021)**. We do not see the reported Thrombomodulin downregulating cytokines TNFα,IL-1α or IL-1β as significant discriminators in any of the models using concentrations, ratios or parameters. We see that CCL-4/MIP-1β $\iff$ Thrombomodulin as a strong discriminator. We do not know if there is a direct interaction between the two analytes, but TNFα has been reported to influence CCL-4/MIP-1β expression **(Ahmad et al. 2019)**. In the model this is an inverse relationship though. Lower CCL-4/MIP-1β $\iff$ Thrombomodulin (ratio<~0.10) is associated with NSTI patients. This equates to higher levels of Thrombomodulin and lower levels of CCL-4/MIP-1β were associated with NSTI patients by the model and the opposite with Non-NSTI patients.

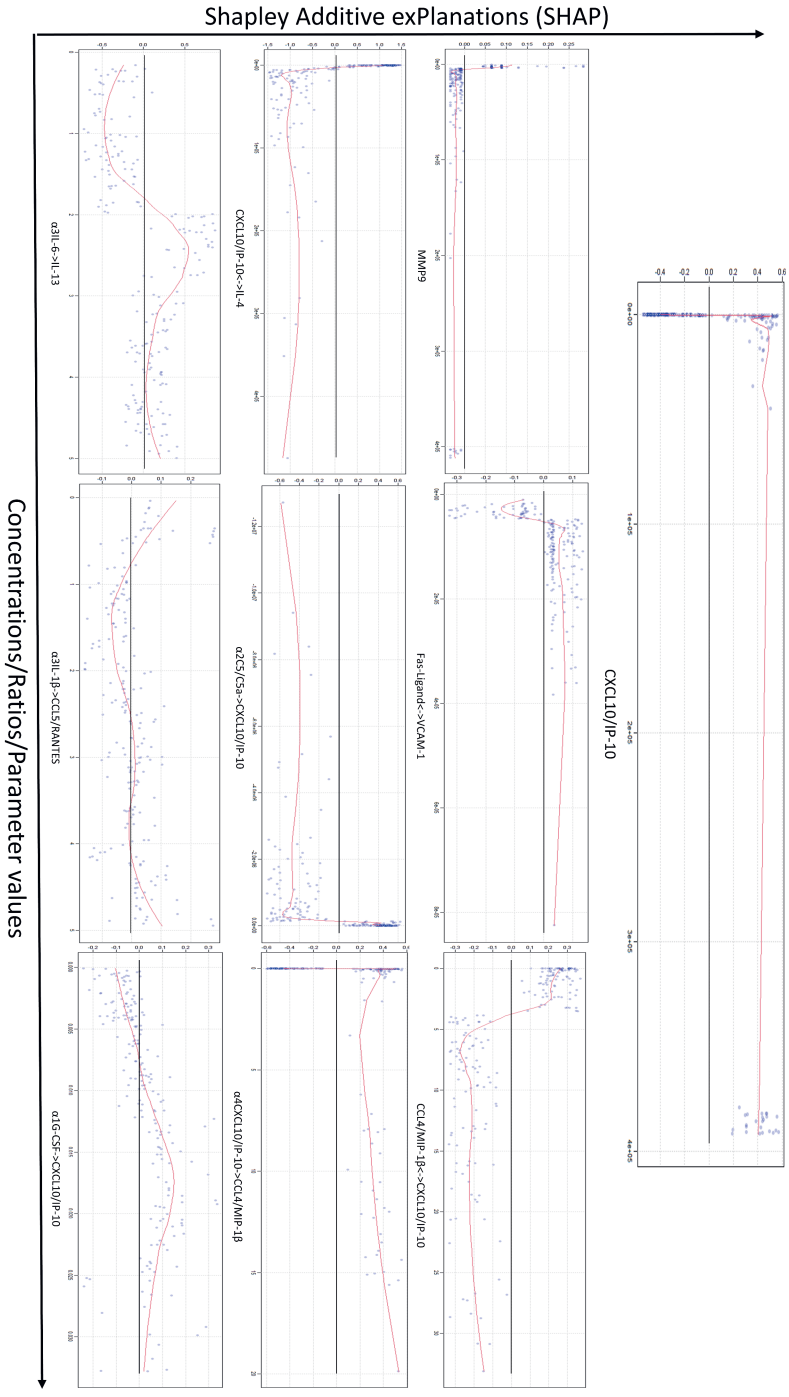

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
α3IL-6->IL-13
CXCL10/IP-10<->IL-4
MMP9
CXCL10/IP-10
α3IL-1β->CCL5/RANTES
α2CS/CSa->CXCL10/IP-10
Fas-Ligand<->VCAM-1
α1G-CSF->CXCL10/IP-10
α4CXCL10/IP-10->CCL4/MIP-1β
CCL4/MIP-1β<->CXCL10/IP-10

Figure 8.11

Figure 8.11: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between mono-microbial NSTI and poly-microbial NSTI. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as mono-microbial NSTI or poly-microbial NSTI. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient as NSTI. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient as poly-microbial NSTI. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

In addition, we find α_2IL-4 $\Rightarrow$ Thrombomodulin and α_4IL-4 $\Rightarrow$ Thrmbomodulin as parameters having the most significant effect on the classification. IL-4 has been experimentally shown to counteract the effects of TNFα and IL-1 induced downregulation of Thrombomodulin using cultured human umbilical vein endothelial cells **(Kapiotis et al. 1991)**. This leads us to hypothesise mechanisms leading to reduction in the suppression of Thrombomodulin particularly in NSTI patients.

8.4.3 Patient outcome

MMP-1, also known as collagenase-1, is an enzyme similar to MMP-8 in many ways. MMP-1 is associated with similar functions as MMP-8 with many more substrates **(Laronha et al. 2020)**. The levels of MMP-1 while classifying based on Mortality is probably indicative of significant tissue damage occurring in the patients. Concentrations of E-Selectin and C5/C5a are also likely indicative of a immune response and don't seem to be very clear discriminators from figure 8.8. C5/C5a has been known to stimulate endothelial cells and induce the expression of E-Selectin and other similar cell adhesion molecules **(Albrecht et al. 2004)**. C5/C5a $\Longleftrightarrow$ MMP-1 also was good at classifying mortality. C5a has been shown to induce the expression of MMP-1 and both C5a and MMP-1 have been shown to be regulated by IL-1 **(Laronha et al. 2020; Speidl et al. 2011)**.

S100A8 is a damage-associated molecular pattern (DAMP) molecule that is released under stress and can act as an activation agent for the recruitment of many immune cells. Molecules of the calgranulin family have been shown to be involved in various inflammatory conditions with their ability to perform surplus cytokine and chemokine-like functions **(S. Wang et al. 2018)**. S100A8 induces the translocation of MyD88 along with the hyperphosphorylation of IRAK-1 and the activation of NF-$\kappa\beta$ during septic shock resulting in elevated expression of TNFα in phagocytes **(Vogl, Tenbrock, et al. 2007)**. A higher concentration of S100A8 (>~900 pg/ml) was associated with the occurrence of septic shock in the Xgboost model. The higher concentration is potentially indicative of higher tissue damage expected in patients with septic shock. S100A8's ability to discriminate the occurrence of septic shock was also reported in **(Medina et al. 2021)**. Concentrations of S100A8, S100A9 and associated complexes have also been associated with higher risks of mortality in septic shock patients **(Dubois, Marcé, et al. 2019)**. Furthermore, concentrations of S100A8 & mono-oxidized S100A8 in plasma were identified by Dubois et al. as being signifi-

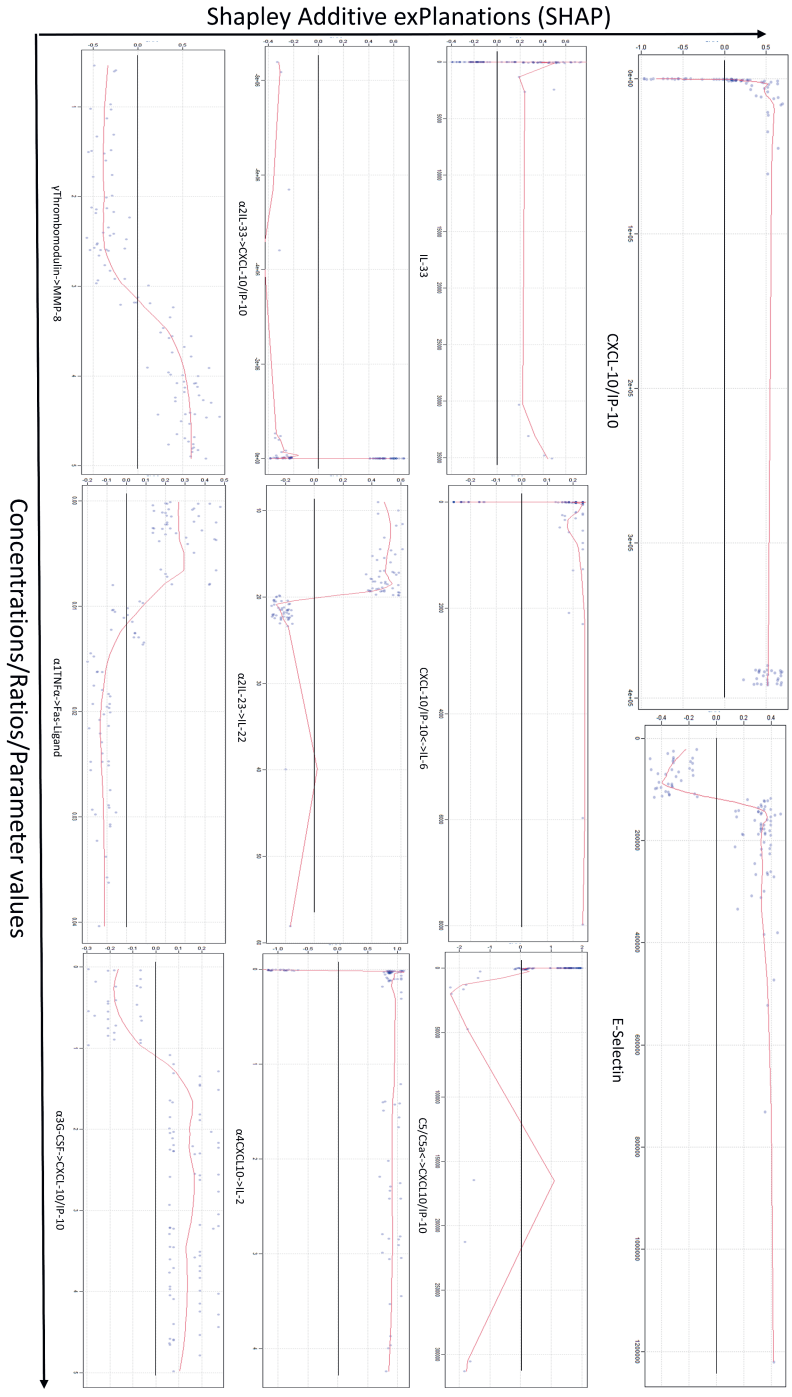

Shapley Additive exPlanations (SHAP)
Concentrations/Ratios/Parameter values
CXCL-10/IP-10
IL-33
α2IL-33->CXCL-10/IP-10
γThrombomodulin->MMP-8
E-Selectin
CXCL-10/IP-10<->IL-6
α2IL-23->IL-22
α1TNFα->Fas-Ligand
C5/C5a<->CXCL10/IP-10
α4CXCL10->IL-2
α3G-CSF->CXCL-10/IP-10

Figure 8.12

Figure 8.12: Select plots showing the ability of analytes, ratios of analytes, and model parameters estimated from analyte concentrations to differentiate between NSTI caused by GAS and NSTI caused by SD. The x-axis denotes either the concentration values in pg/ml, ratios or the parameter values. The y-axis shows the effect on the model classifying a sample as NSTI caused by GAS or NSTI caused by SD. The dark line marks the 0-value on the y-axis. Each blue dot represents a sample/patient. The further above the black line, a sample is on the y-axis (positive values), the higher influence that sample with the concentration/ratio/parameter value has on the model classifying that patient with NSTI caused by GAS. Similarly, the further below the dark line, the samples are, the higher influence the sample has on classifying the patient with NSTI caused by SD. The red line shows the curve fit by the model on the particular data and sheds light on the model's classifications process.

cantly increased in non-survivor septic shock patients **(Dubois, Payen, et al. 2020)**. IL-13 was reported to suppress S100A8 expression in human keratinocyte cell line when studying inflammatory skin diseases. In our model, we find lower values of ($<\sim 50$) α_4IL-13 $\Rightarrow$ S100A8 to be associated with septic shock occurrence. It is interesting to note, that unlike S100A8, lower concentrations of C5/C5a ($<\sim 1250$ pg/ml) were associated by the model with septic shock occurrence. The discriminatory ability of C5/C5a $\iff$ IL-6 correlates with the study showing the regulation of C5a during sepsis by the production of IL-6 in rodents linking it to multi-organ failures **(Riedemann et al. 2004)**.

8.4.4 Microbial aetiology

When we used the models to classify patients based on the microbial aetiology, we found CXCL-10/IP-10 to have a good discriminatory ability to distinguish between mono- and poly-microbial infections as well as Group A streptococcal infections and other streptococcal infections. Thanert et al. found a set of genes encoding interferon-inducible mediators including CXCL-10/IP-10 to be highly expressed in patients with mono-microbial infections and particularly when the infection was caused by S.pyogenes **(Thänert et al. 2019)**. Low concentrations of CXCL-10/IP-10 in the Xgboost models were associated with mono-microbial infections and infections caused by S.pyogenes. S.pyogenes M1 serotype was found to secrete the protein streptococcal inhibitor of complement (SIC) inhibiting the activity of chemokines **(Egesten et al. 2007)**. In a previous study analysing host-pathogen gene associations CXCL5 and CXCL9 genes were found to be deferentially associated between mono- and ploy-microbial infections where S.pyogenes was the causal organism **(Jahagirdar, L. Morris, et al. 2022)**. CXCL-10/IP-10 along with IL-2 was reported as a biomarker for NSTI-type identification by LP Medina et al. and it was found to be significantly associated with IL-2 which was associated with IL-4 in their network analysis **(Medina et al. 2021)**. IL-2 and IL-4 were also found to have a high AUC in their ROC analysis for NSTI-type identification. CXCL-10/IP-10 was also negatively associated with G-CSF in mono-microbial patients with septic shock **(Medina et al. 2021)**. We find α_4CXCL-10/IP-10 $\Rightarrow$ IL-2, CXCL-10/IP-10 $\iff$ IL-4 and α_4G-CSF $\Rightarrow$ CXCL-10/IP-10 to significantly distinguish between GAS & SD and mono- & poly-microbial infections respectively.

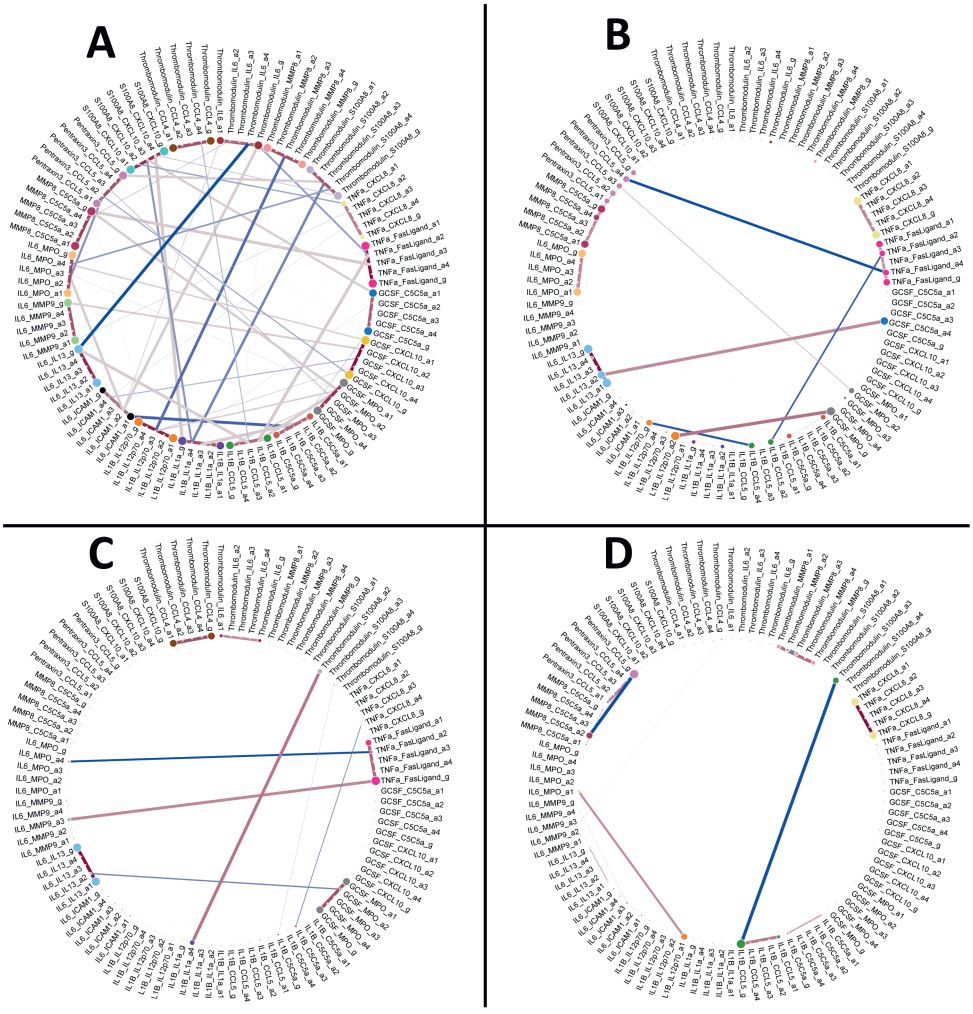


Figure 8.13: The associations between the estimated parameter values in the form of networks. Each dot in the figures, also known as a node, represents a parameter that was estimated. Different colours for the nodes represent different analyte pairs as pro- and anti-inflammatory analyte. The links joining the nodes, also known as edges depict the existence of an association between the two parameters. Edges on the colour gradient of red showcase a positive association and the edges on the colour gradient of blue showcase a negative association. The positive association represents the increase or decrease of the value of one parameter on the increase or decrease of the value of the other parameter. The negetive association symbolises the increase or decrease of the value of one parameter on the decrease or increase of the value of another parameter. The edges represent two-fold values of associations. They show the association of one parameter with another when the effect of all the other parameters are removed and they also represent the probability of the existence of the associations. Edges below the probability threshold (0.95) are not shown. Network in sub-plot (A) is built from parameters estimated from analytes measured in NSTI patients. Network in sub-plot (B) represents non-NST patients. Network in sub-plot (C) represents cellulitis patients and sub-plot (D) represents surgical controls.

8.4.5 Limitations of the study

This study is an exploratory probe to extrapolate information regarding the underlying mechanisms and generate hypotheses using existing data. Limitations in this study come from both the data samples and computational methods perspective. Limitations include relatively small number of samples, heterogeneity of the patients with respect to comorbidities, time of infection and admission to the hospital, and lack of dynamic analyte concentration data due to both mechanistic and ethical constrains. Computationally, limitations include the many assumptions we make, the potential of characterising the parameter estimation as extreme data transformation, the pro- and anti- inflammatory functions not being foregone conclusions, and the fact that not all analytes in this study mimic the proposed cytokine behavior modelled in the baker model. However, clinicians using standardised SOPs, a similar prospective observational study design in all clinical centres, a thorough exploration of the model behaviour, stringent evaluation methods for model performance and the use of stringent statistical approaches reinforces the study design.

8.5 Conclusions

In this study, we conduct an exploratory computational experiment aimed at enhancing our understanding of the underlying mechanisms in NSTI. By generating data-driven hypotheses focused on the pairwise relationships between cytokines and other plasma analytes involved in the human immune response, we delineate ratios and parameters that embody the pro- and anti-inflammatory characteristics of cytokines. We demonstrate the potential of these ratios and parameters to differentiate between NSTI and controls, distinguish microbial aetiologies, and predict patient outcomes. This research deepens our comprehension of key mechanisms in NSTI and can potentially help in the development of more effective diagnostic and prognostic strategies for NSTI.

8.6 Funding

This work was supported by the Center for Innovative Medicine (CIMED) and Region Stockholm (no. 20180058); the Swedish Research Council (2018-02475); the European Union Seventh Framework Programme (FP7/2007-2013) under the grant agreement 305340 (INFECT project); the Swedish Governmental Agency for Innovation Systems (VINNOVA), Innovation Fund Denmark (8114-000005B) and the Research Council of Norway under the frame of NordForsk (project no. 90456, PerAID); the Swedish Research Council, Innovation Fund Denmark (8113-000009B), the Research Council of Norway, the Netherlands Organisation for Health Research and Development (ZonMW) and DLR Federal Ministry of Education and Research, under the frame of ERA PerMed (project 2018-151, PerMIT); and the Swedish Children's Cancer Foundation (TJ2018-0128).

8.7 Author contributions

SJ and ES designed the study; SJ, ML, CK, OBC performed the analysis; SJ, LPM, KAM, MS, OH, ANT, VAdP, and ES provided advice and interpretation of results; SJ visualised the results; SJ wrote the manuscript; SJ, LPM, KAT, OH, SS, VAdP, and ES critical revision. All authors reviewed the manuscript.

Chapter 9 Chapter

Sanjeevan Jahagirdar[1*], Shruti Setty[20*], Zoë Robaey[21], Vitor A. P. Martins dos Santos[7,8]
***Contributed equally**

Turn to page 377 for author affiliations

This chapter is prepared for publication

Wisdom of the Informed Crowds

Abstract

The solutions to complex scientific and societal problems often require multi-/inter-/trans-disciplinary approaches. The responsibility of finding technical solutions to multifaceted problems often falls on the shoulders of early career researchers. However, the perspectives of these early career researchers are often underappreciated. We present a creative problem-solving model for technical problems in inter-disciplinary environments addressing managerial provisions, objectives, type of environment, technical deliberations, and sceptical discourse from the perspective of early career researchers in the computational and mathematical sciences collaborating with experts in other fields. We also call to attention the existence of a mathematical linguistic barrier and the need for linking different mathematical formulations.

9.1 Background

The heightened focus on interdisciplinary collaboration has been intensifying in recent decades, making multi-/inter-/trans-disciplinary projects a staple in contemporary scientific endeavours **(Med 2006; Fauchald et al. 2005)**. The resolution of complex societal and scientific issues often necessitates a multidimensional approach, thus underscoring the inherent value and advantages of integrating expertise across different disciplines **(Med 2006; Choi and Pak 2007; Whitfield et al. 2004)**. Such a synergistic approach frequently results in assembling investigators from diverse domains, thereby fostering the emergence of fresh perspectives, spurring innovative methodologies, catalysing the creation of novel paradigms and unique frameworks, and ultimately engendering new languages within scientific (and eventually societal) discourse **(Weaver 2008)**.

The shift towards interdisciplinary collaboration is notably reflected in the tendencies of funding agencies to favour research proposals involving various disciplines, as well as academic institutions' initiatives to establish departments and curricula embodying a multidisciplinary ethos **(Med 2006)**. Senior scientists, invested in long-term visions, and often involved in securing said grants, can be unaware of the technical concerns while problem-solving for solutions addressing such multifaceted issues. Despite the proposal of numerous problem-solving models and strategies focusing on managerial aspects to nurture a multidisciplinary approach, the onus of technical problem-solving frequently resides with early career researchers **(MacLeod 2018; O'Loughlin et al. 1999)**. However, the perspectives of these early career researchers remain critically underrepresented and underreported in the current discourse **(Pannell et al. 2019)**.

In the context of problem-solving, the perspectives and tasks of early career researchers and senior scientists can vary notably. Early career researchers can carry an explorative mindset while being comfortable with the latest techniques and trends and can approach problems with a keen openness to novel ideas and unconventional methods. Meanwhile, senior scientists, backed by their historical contexts and knowledge focus more on the fundamental concepts with traditional approaches. They are able to integrate findings and interpretations across a significantly wider span of disciplines. They tend to focus on long-term advancements and developments and align the research questions with long-standing paradigms. Both these perspectives are incredibly important in scientific research and in implementing innovative and comprehensive solutions to difficult problems. In this chapter, we solely focus on the perspectives of the early career researcher, especially from the perspective of early career researchers in the domain of computational and mathematical sciences collaborating with experts from other scientific fields.

9.1.1 Distinction in multi-/inter-/trans-disciplinary projects

In the context of scientific collaboration and scientific problem-solving the terms multi-disciplinary, inter-disciplinary and trans-disciplinary are used to describe distinct multiple disciplinary approaches to varying degrees albeit on the same continuum **(Choi and Pak 2006)**. Definitions put forth by Choi et al. suggest that multi-

disciplinary research draws on knowledge from different disciplines but stays within the boundaries of those fields, meanwhile, Interdisciplinary research analyses, synthesises and harmonises connections between disciplines into a coordinated and coherent whole. Finally, trans-disciplinary research integrates the natural, social and health sciences in a humanities context, and in so doing transcends each of their traditional boundaries **(Choi and Pak 2006)**. As these formulations fall on a continuum, so do some of the issues (while problem-solving) that occur in these environments concur. We aim to address those specific issues in this manuscript.

9.1.2 Scientific problems can be wicked

Problems can manifest themselves in a myriad of forms, including but not limited to puzzles, algorithmic quandaries, narrative complications, decision-making conundrums, troubleshooting tasks, diagnostic challenges, and design issues. Yet, there is a prevailing trend categorising problems into structured, unstructured, and ill-structured types, contingent upon the complexity and clarity of the problem definition and the solution methodology **(Reed 2016)**. The scientific problems often confronted in multi-/inter-/trans-disciplinary endeavours frequently align with ill-structured problems or could even be classified as wicked problems, characterised by an absence of definitive formulation, open-ended causal chains, and solutions that resist binary true-false classifications **(Lawrence et al. 2022)**. The designs of problem-solving models often aim to establish a streamlined, linear process to expedite the discovery of a solution **(Chakravorty et al. 2008)**. However, this can inadvertently stifle creative cognition as a preference towards reducing uncertainty has shown to induce a bias against creative solutions **(Cassotti et al. 2016)**. The intricacy of these tasks is further amplified when experts from varied disciplines and backgrounds use distinct languages, encompassing unique jargons and acronyms **(Choi, Pang, et al. 2005)**. The proposed resolutions for these multifaceted issues often adopt a top-down, heuristic disposition, which may unintentionally curtail bottom-up creativity **(Fedor-Freybergh 1999; H. A. White 2020)**.

Engaging in critical and sceptical discourse constitutes a fundamental part of discovering objective scientific solutions and is well established in the overall scientific process **(Osborne 2010)**. However, it is essential to recognise that creative problem-solving is executed not by objective machines, but by teams of humans. These subjective individuals inherently bring their biases, principles, experiences, ambitions, long-term goals, passions, and emotions to the process **(Oeberst et al. 2023; Gesiarz et al. 2019; Pannucci et al. 2010)**. Here, we keenly recognise that both convergent and divergent thought processes have a key role to play in the scientific problem-solving process, however, the requirements to conduct these thought processes are very different. In this article, we propose a model for creative problem-solving, specifically tailored to address technical issues encountered in inter-disciplinary projects that are not always helped by the reliance on the heuristic knowledge-based approaches of individual fields. Furthermore, we draw attention to the presence of a mathematical-linguistic barrier inherent in these projects and offer potentially beneficial strategies borrowed from the domain of software engineering. The intention behind these strategies is to open a dialogue that may gradually reduce this barrier in a bottom-up fashion, fostering more effective cross-disciplinary communication

and collaboration.

9.2 Problem solving in inter-disciplinary projects

9.2.1 Context

Here, we introduce a model (Figure 9.1) for creative problem-solving, specifically tailored for technical challenges encountered in inter-disciplinary settings. This model has been developed based on the collective insights and experiences of two early-career researchers actively engaged in inter-disciplinary projects. One of the researchers is involved in a systems medicine project focusing on infectious diseases, collaborating with medical doctors, microbiologists, and systems biologists. The other researcher is engaged in a paleo-climate modelling project focusing on modelling the climate system, working in conjunction with earth scientists, ecologists, and computational modellers.

While there are models that approach these complex projects from a managerial perspective, outlining actions and leadership styles for project managers, the model we present here is distinct in its focus. It aims to address the needs of early career scientists and emphasises the technical steps necessary for identifying optimal technical solutions, a responsibility that often falls on the shoulders of early career researchers. This model strives to distinguish between the convergent thought processes and critical or sceptical discourse, and the creative cognition process necessary for generating unique solutions in the steps taken from information gathering to the evaluation of proposed solutions. Furthermore, the focus of this model is not just on technical steps but the environment necessary to be able to problem-solve in such multi-faceted projects.

The model consists of an outer ring defining the general steps for solving problems and six inner layers focused on the managerial provisions, objectives, environment within the team, technical details, convergent, and critical discourse.

9.2.2 The modified 5S model

The outer ring of the model depicted in figure 9.1, illustrated in varying shades of purple, is a modified rendition of the five-step (5S) model initially proposed by **(Chakravorty et al. 2008)**. The original model delineates the following steps to expedite the discovery of a solution: (1) Problem identification, (2) Information gathering, (3) Generation of alternative solutions, (4) Evaluation of solutions, and (5) Implementation of the optimal solution(s) **(Chakravorty et al. 2008)**. This model is similar to many sequential problem-solving models that have been described consisting of four to six sequential steps. However, in the context of scientific research, this model forces a premature convergence by implementing the best solution before gathering alternate solutions and does not incorporate steps for collaborative brainstorming. Additionally, the sequential flow may not be conducive towards the dynamic nature of scientific research and identifying the underlying causes of the problem.

Our adaptation of this model shifts the focus from expediting the problem-solving process to identifying the best solution within the available time frame. This modification is based on evidence suggesting that rushing the problem-solving process

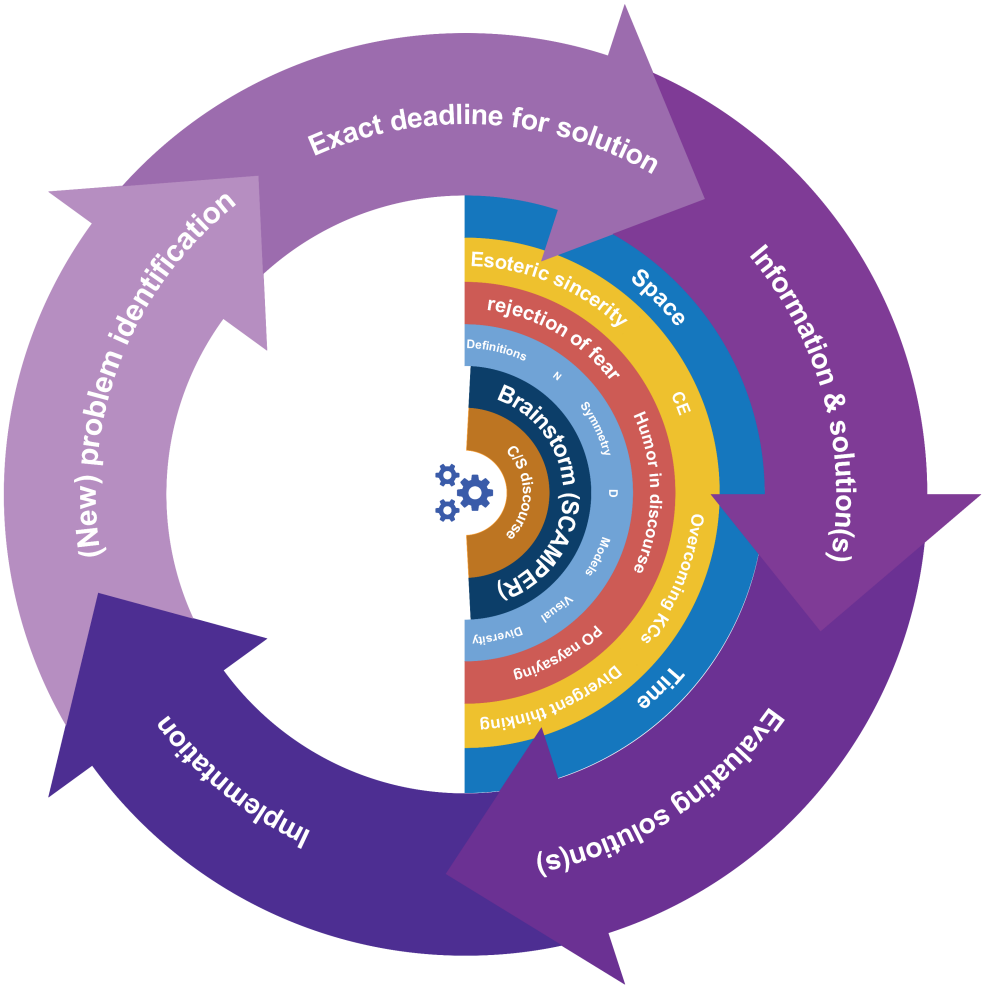


Figure 9.1: Creative problem-solving model for technical problems in inter-disciplinary projects. **CE: Conceptual expansion, KCs: Knowledge constraints, PO: Postponement of, N: Nomenclature, D: Descriptions, C/S: Critical/Sceptical.** The outer ring in shades of purple is a modification of the 5s model introduced in **(Chakravorty et al. 2008)**. Here, our model focuses on the right side of the ring from information gathering to the determination of the best solution. The model is divided into layers separating divergent and convergent thought processes for the team as well as separating provisional requirements of the early career researchers from management and the technical process of solving problems. The first layer in blue focuses on the most important managerial provision required by the early career researcher. The second layer in yellow focuses on the objectives to focus on for the problem-solving exercise, which will eventually lead to a good solution in a multi-disciplinary environment. The third layer in red focuses on the environment that is established for solving complex problems. This is the responsibility of both early career researchers and managerial staff. The fourth layer in light blue focuses on the technical aspects towards finding a solution. The fifth layer in dark blue, starts to shift the attention from divergent thought processes to convergent thought processes and brings together different ideas to move the team towards solution(s). The sixth layer in ochre focuses on the critical and sceptical discourse that is required for objective scientific solutions, separating it from the playful creative process.

can undermine creative cognitive processes **(Cassotti et al. 2016)**. Consequently, we incorporate the time element as an external factor, which is determined concurrently with the problem identification process.

Therefore, our revised 5S model comprises the following steps: (1) (New) Problem identification/formulation, (2) Establishment of a precise deadline for the required solution, (3) Information gathering and generation of alternative solution(s), (4) Evaluation of solutions, and (5) Implementation of the optimal solution(s). This approach emphasises the importance of identifying the best solution within the available time frame and maximising the use of this time. Furthermore, we make the model into a circular model rather than a sequential one based on our experience with scientific problem-solving.

9.2.3 Managerial provisions

The first layer of the model, depicted in blue, pertains to managerial considerations. Rather than addressing the needs from the perspective of large-scale project management or leadership styles, this layer concentrates on the requirements of early career researchers tasked with devising technical solutions to complex, multifaceted problems.

The concept of interdisciplinary problem-solving has been metaphorically likened to "bricolage" by Andrade et al. **(Andrade et al. 2014)**. In the realm of the arts, "bricolage" refers to the creation of a new work from a diverse array of available components. In the context of our model, we propose that early career researchers necessitate an unrestricted or safe space, coupled with a defined time period and patience.

The term safe space here does not refer to physical safety of a group as studied in normal discourse **(Flensner et al. 2019)**, but rather to the concept proposed by Du Preez, where it signifies a place conducive to taking intellectual risks **(Du Preez 2012; Andrade et al. 2014)**. The process of problem-solving in interdisciplinary environments often demands stepping beyond the heuristic methods of individual disciplines and necessitating intellectual risk-taking **(Andrade et al. 2014; Graff 2016)**.

A dedicated time period for fostering interdisciplinary processes and relationships among team members is crucial. Early career researchers require time to familiarise themselves with the assumptions and technical languages of other disciplines, navigate through frustrations, revisit ideas, and build trust **(Andrade et al. 2014)**. In this context, Andrade et al. emphasise that patience is paramount and caution against tight, rigid, and inflexible schedules for interdisciplinary problem-solving **(Andrade et al. 2014)**.

9.2.4 Objectives en-route to the solution

The model's second layer, represented by the colour yellow, deliberates on objectives capable of facilitating effective interdisciplinary solutions. The emphasis should be on considering a wide range of objectives, including but not limited to technical skills, those grounded in systems thinking, and cognitive-based objectives during the problem-solving process **(Barber 2018; Blanco et al. 2019)**. Here, we highlight 4 objectives we think are crucial: (a) Esoteric sincerity, (b) Conceptual expansion,

(c) Overcoming domain-specific knowledge constraints, and (d) divergent thought processes.

In conjunction with objectives designed to help creative cognitive process, it's equally critical to factor in value-based considerations in the problem-solving process **(Škėrienė et al. 2020)**. An approach to problem-solving, firmly rooted in a shared value system, significantly contributes to building trust and enhancing the validity of the outcomes within an interdisciplinary team **(Škėrienė et al. 2020)**.

In this context, we propose fostering an ethos of esoteric sincerity within the team. Here, "esoteric" does not imply an exclusionary or elitist approach to knowledge, accessible only to a select few. Rather, it designates the fact that the depth and specificity of information from different domains that only certain team members, due to their specialised expertise, can fully comprehend. An objective towards sincerity is indispensable to achieving truly reliable and innovative solutions to interdisciplinary challenges.

Conceptual expansion is a critical facet of the creative cognition required for addressing intricate, multifaceted problems **(A. Abraham et al. 2012)**. Albeit obvious, it is important here to emphasise that the majority of creative problem-solving tasks do not conform to a binary solution framework (yes/no, black/white answers), and can necessitate the use of drawing or verbal articulation, also potentially stimulating brain activity related to movements **(Carlsson et al. 2000; H. A. White 2020)**. Given the unpredictable nature of creative problem-solving, both in terms of its voluntary initiation and the difficulty in knowing the precise moment of creative response generation **(Fink et al. 2009)**, we suggest the establishment of a team objective aimed at fostering conceptual expansion in the interdisciplinary domain.

The bulk of scientific research adheres to the conventions established within specific scientific domains **(MacLeod 2018)**. These domain-specific structures, while essential, can inadvertently hinder problem-solving in interdisciplinary settings due to inherent knowledge constraints **(MacLeod 2018; Garriga et al. 2013)**. These "conventions" often arise through a rigorous discovery process itself. Breaking the boundaries of existing conventions leads to new conventions. Thus, there is always an intrinsic stress between groundbreaking and established frameworks. In light of this, we advocate for an interdisciplinary team objective aimed at surmounting these knowledge constraints by minimising restrictions to specific domain-external knowledge.

Divergent thinking serves as an apt metaphor for the cognitive processes that can generate original ideas **(Runco et al. 2012)**. The importance of setting an objective to foster divergent thinking within a team for interdisciplinary problem-solving cannot be overstated **(Basadur and Hausdorf 1996)**. In this context, we reference the four attitudes towards divergent thinking as proposed by Basadur et al., which are (a) preference for generatio of ideas, (b) a tendency to not make premature critical evaluations of ideas, (c) valuing new ideas, and (d) a belief that critical thinking is not bizarre **(Basadur, Finkbeiner, et al. 1983; Basadur and Finkbeiner 1985)**. However this concept is not limited to these aforementioned four attitudes.

9.2.5 Environment within the team

The third layer of the model, denoted by the colour red, emphasises the environment necessary to enable interdisciplinary problem-solving. This environment pertains to

the needs of a team of early career researchers tasked with solving complex multi-faceted problems. Both managers and early-career researchers share the responsibility of cultivating this conducive environment, which is integral to interdisciplinary problem-solving. We present here three elements that should be inherent in this environment: (a) a rejection of fear of failure, (b) the incorporation of humour in discourse, and (c) the deferment/postponing of naysaying.

The generation of novel ideas through creative cognitive processes is integral to creative problem-solving in interdisciplinary environments. However, it is crucial to acknowledge that despite its endorsements, creativity is an inherently risky behaviour **(Tyagi et al. 2017; Y. S. Lee et al. 2017)**. Studies have demonstrated that the apprehension of negative feedback significantly deters individuals from sharing creative ideas **(E. W. Morrison 2011; Y. S. Lee et al. 2017)**. Merely encouraging early career researchers to propose creative ideas is of limited value without considering the managerial response to those ideas. Lee et al. reveal an implicit bias against creative solutions when managerial appraisal is geared towards reducing uncertainty **(Y. S. Lee et al. 2017)**. Moreover, research involving classroom students has indicated a decline in engagement when students harbour a fear of failure **(Nakhla 2019)**. Studies have also found that a high implicit fear of failure leads to slower information assimilation and diminished learning rates during a task, with the correlation increasing under conditions of frustration **(Lerche et al. 2018)**. Therefore, the environment in which early career researchers seek solutions to interdisciplinary problems must actively reject the fear of failure, not just by means of words, but also considering the effect of actions and policies.

Humour, a prevalent component of human interactions, can hold considerable value for early career researchers in an interdisciplinary environment, particularly in terms of stress reduction, the enhancement of group cohesion, communication, creativity, and the cultivation of a positive organisational culture **(Romero et al. 2006)**. It further exhibits the capacity to alleviate tensions that may arise from different domain-specific interpretations of concepts **(Moats 2021)**. Evidence suggests that the inclusion of humour in communications enhances comprehension and recall, as compared to communications devoid of humour **(Schmidt 2002; Schmidt and A. R. Williams 2001; Wood et al. 2011)**. Additionally, humour fosters a more relaxed, playful, and secure environment, promoting free association and facilitating more risky behaviours essential in problem-solving endeavours **(Consalvo 1989)**.

The proposition of the postponement of naysaying not only reinforces the second attitude of divergent thinking as outlined by Basadur et al. (the avoidance of premature critical evaluations of ideas) **(Basadur, Finkbeiner, et al. 1983; Basadur and Finkbeiner 1985)**, but also aligns with what can be described as "insight problems" **(Bagassi et al. 2020)**. As Bagassi et al. elucidate, it is often necessary to abandon a domain-specific interpretation (fixation) to reorient one's thinking towards an interpretation more pertinent to the interdisciplinary solution (aim of the problem-solving exercise) **(Bagassi et al. 2020)**. Failure to do so can result in misinterpretations and misunderstandings of crucial aspects of the interdisciplinary solution **(Bagassi et al. 2020)**. Therefore, a deliberate postponement of naysaying is an essential component of the conducive environment.

9.2.6 Technical deliberations

The fourth layer, indicated by a light blue hue, underscores the vital technical deliberations and considerations that can be particularly beneficial for early career researchers navigating multifaceted problems within an interdisciplinary settings. Within this context, we propose several practices potentially conducive to successful outcomes in such an interdisciplinary milieu: (a) a precise definition of all the constituent elements and their interactions, (b) establishing a common nomenclature in the interdisciplinary environment, (c) leveraging mathematical, theoretical, and metaphorical symmetries, (d) understanding the definitions of all the elements from the perspective of all the different domains involved in the interdisciplinary project, (e) data-driven, unbiased visualisations, (f) exploratory modelling, and (g) considerations of diverse inputs.

The precise identification of all constituent elements and their interactions is vital for achieving a comprehensive understanding of the system at large. In an interdisciplinary framework, these elements may derive from various disciplines and may be entrenched in the inherent assumptions of each discipline's technical language **(Choi, Pang, et al. 2005; Choi and Pak 2007)**. Nevertheless, their exact definition is pivotal within the context of interdisciplinary research as these elements could foster knowledge sharing at certain levels, even amidst seemingly disparate disciplines **(Choi and Anita 2008)**.

The hurdles presented by the difference in language across different domains can only be overcome through time and patience. Nonetheless, investing in a **consistent nomenclature**, specifically tailored to articulate the elements and their interactions in the interdisciplinary problem, is a worthwhile effort. Monteiro et al. found that matching disparate domain representations could be an effective strategy to facilitate cognitive transmission in interdisciplinary collaborations **(Monteiro et al. 2009)**.

Such nomenclature, derived collaboratively from the associated domains, can serve as a bridge connecting the language systems and alleviating communication gaps. This endeavour can harmoniously coincide with the detailed description of all elements, features, and their interrelations from various domain perspectives. This practice facilitates the cultivation of an interdisciplinary comprehension of the problems at hand.

The concept of **symmetry** has been recognised as a significant tool, wielding substantial influence in mathematical problem-solving approaches **(Leikin et al. 2000)**. In the realm of fundamental physics, symmetry underscores the consistency and invariance principles intrinsic to the laws postulated in this field **(Gross 1996)**. Such symmetry need not be strictly mathematical and can be exploited to formulate inferences and assumptions about intricate problems, which can be supplemented with playful experimental and model-based testing.

In an interdisciplinary context, symmetries, whether they are mathematical, theoretical, or metaphorical, can be utilised to infer that an observation made in one scenario may recur in another analogous instance. Symmetries can further be incorporated into visualisations to enable a system-wide understanding.

Visualisation itself stands as a potent tool, serving not merely as a means of communication but also as an instrument to discern the relationships among various inter-disciplinary concepts and entities **(A. Jacobs et al. 2008)**. These visualisation tools not only exhibit potential for reapplication across a multitude of interdisci-

plinary problems but can also be adopted by researchers external to the original collaboration for conveying similar concepts **(Kirby et al. 2013)**.

An effective visualisation possesses the ability to elucidate the relationships among multivariate datasets, provide a system-oriented perspective, and offer aid in choosing a path towards decision-making, all while conveying essential information **(Borkin et al. 2013; Dyson 2016)**. It is paramount that visualisations are human-readable, data-driven, and unbiased, as these characteristics underpin all of the assumptions discussed above. This tenet is followed up in **Chapter 10** *Automated NETwork VISualisation with anvis* where we develop a visualisation tool for network models.

Modelling systems can significantly enrich the problem-solving process within interdisciplinary contexts. Predictive and exploratory models can yield insightful information pertaining to system behaviours. Employing a playful, exploratory modelling methodology can enable the testing of assumptions, the generation of hypotheses, and the creation of data-informed perspectives.

In the realm of interdisciplinary problem-solving, the necessity for the inclusion and consideration of **diverse inputs** cannot be overstated. A substantial body of evidence suggests that heterogeneous teams, composed of varying types of thinkers, demonstrate superior performance over homogeneous groups in complex tasks, yielding enhanced problem-solving capacity, improved innovation, and more precise predictions, thus delivering superior results **(Swartz et al. 2019; Page 2019; Freeman et al. 2014; AlShebli et al. 2018)**.

The onus for this diversity is multilayered within this model; assembling a diverse team with equal opportunities is the manager's responsibility (not the focus of the model), while creating an environment that fosters varied thought processes, echoing the concepts discussed in "postponement of naysaying", is incumbent upon all team members. Early career researchers also bear the responsibility of considering technical inputs from diverse thought processes. It is also important to refrain from hoarding dismissive attitudes towards unanticipated or previously unconsidered methodologies with statements like "this is weird", in order to sustain this environment within an interdisciplinary team.

Diversity can encompass a myriad of aspects, including academic background, age, gender, sexual orientation, ethnicity, nationality, culture, religion, geography, disability, socioeconomic status, area of expertise, level of experience, neurodiversity, thinking styles, and non-academic skill sets among others **(Swartz et al. 2019)**. A strong correlation has been observed between the presence of ethnic diversity among authors of scientific publications and the magnitude of the scientific impact **(AlShebli et al. 2018)**. People's varied experiences can propel unique perspectives and significantly augment the innovative potential of an interdisciplinary team.

9.2.7 Convergence of ideas and critical outlook

In this model, we endorse for a separation of divergent and convergent thought processes, not the elimination of convergent thought processes. Creative thinking could be argued to be a two-part equation, comprising both the generation and the evaluation of novel concepts **(Cropley 2006)**. The initial layers of this model concentrate on the creation of novelty. Once armed with a database of novel ideas, the fifth and sixth layers, denoted by the colours dark blue and ochre respectively, focus on

the critical appraisal of these novel ideas, employing convergent thought processes and a critical/sceptical discourse. In these discursive contexts, the importance of pre-existing knowledge and domain-specific comprehension cannot be overstated **(Cropley 2006)**.

We put forth two evaluative steps designed to identify optimal solutions from the array of novel ideas: (a) Brainstorming sessions with experienced domain experts can amalgamate innovative ideas with existing knowledge, thereby preventing pseudo-creativity and further enhancing the ideas, and (b) Critical/Sceptical discourse with senior researchers, managers, domain experts, and early career researchers can scrutinise every facet of the proposed ideas, as is a common practice in scientific discourse.

Various brainstorming techniques have already been proposed, including SCAMPER, random connection, schema violation, and simple ideation, any of which, or a combination thereof, could be deployed for brainstorming **(X. Gu et al. 2022; Ozyaprak 2016)**. The SCAMPER technique, in particular, could prove beneficial in this context, as it employs directed, idea-stimulating questions to suggest expansions or alterations of existing ideas **(Eberle 1972)**. The acronym SCAMPER stands for Substitute, Combine, Adapt, Magnify/Modify, Put to other uses, Eliminate, and Rearrange/Reverse **(Eberle 1972; Serrat et al. 2017)**. For a more comprehensive understanding of this technique, we refer interested readers elsewhere **(Serrat et al. 2017)**.

9.3 Reducing the mathematical linguistic barrier in communication

The evolution of technical languages is heavily influenced by domain-specific necessities, leading to diverse linguistic usages among professionals from varying fields **(Choi, Pang, et al. 2005)**. Consequently, identical terminology may convey differing interpretations across distinct disciplines **(Choi and Pak 2007)**. In several scientific arenas, the reliance on quantitative measurements makes mathematics an essential component of their communicative language. Nevertheless, mathematical language does not present itself as a monolithic entity, but rather as an assemblage of numerous descriptive systems. Moreover, a considerable portion of mathematical regulations are arbitrary assignments rooted in historical precedents **(Sfard 2007)**. Despite this randomness, the universal agreement and habitual usage by numerous individuals contribute to their meaning. Interestingly, the translation between these descriptive systems often yields innovative and beneficial outcomes **(Galperin 2003)**. Nevertheless, this mode of mathematical discourse further exaggerates the linguistic hindrances in communication. Calls have been made to consolidate some of these mathematical paradigms **(Lasenby et al. 2000)**, but these are typically top-down approaches that do not address the unique needs of interdisciplinary contexts.

Propositions have already been put forth to enhance the accessibility, comprehensibility, and reusability of data storage/collection practices and research softwares **(Wilkinson et al. 2016; Barker et al. 2022)**. Here, we aim to initiate a discussion on mathematical formulations, and the need for a similar conversation to establish similar recommendations for mathematical formulations. The linking of different mathematical formulations with provenance concepts, research softwares, analogous

concepts interpreted through distinct metaphors, and analogous concepts transposed into different descriptive systems, along with a standardisation of nomenclature when communicating mathematical equations, could offer several benefits. These benefits, applicable to both domain (mathematicians) experts and non-domain experts, could enhance decision-making in method selection, inter-disciplinary communication, and visibility of novel methodologies. Crafting explicit recommendations for realising these goals requires input from a variety of fields and experts, which is out of the scope of this paper. However, our intention is to open this critical conversation, that can be pursued and explored in further discussions.

9.4 Conclusions

The number of inter-disciplinary projects have been increasing in science to solve complex multifaceted problems. We propose a creative-problem solving model for technical problems in inter-disciplinary projects from the perspective of early career researchers addressing (a) managerial provisions required by early career researchers, (b) objectives to focus on during the problem-solving exercise, (c) establishing an environment for complex problem-solving, (d) technical aspects that can help while finding a solution, (e) Convergent thought processes and (f) critical and sceptical discourse. We end this paper with starting a discussion on the existence of mathematical linguistic barriers and the need to establish bottom-up recommendations for linking different mathematical formulations, translations and research softwares.

Chapter 10

Chapter

Robert Koetsier[1,22§], Shruti Setty[20], Edoardo Saccenti[1*], Ingrid van de Leemput[20*], Maria Suarez Diez[1*], Vitor A. P. Martins dos Santos[7,8*], Sanjeevan Jahagirdar[1]
***These authors are listed in alphabetical order.**

Turn to page 377 for author affiliations

This chapter is prepared for publication

Automated NETwork VISualisation with *anvis*

Abstract

Network models offer a flexible mechanism for characterising complex systems based on their constituent elements and inter-element interactions. These models can be effectively visualised as graphs or node-linked diagrams, facilitating system-level understanding and decision-making. However, as the number of nodes and edges within a network escalates, the resulting visualisations may fail to convey meaningful information, necessitating laborious manual refinement to create an informative network topology and as a consequence introducing user bias. In response to this challenge, we introduce an Automated Network VISualisation package (anvis). anvis automates the selection of variables that define the network topology, enabling the creation of multiple data-driven, publication-ready visuals that are easily interpretable. These visuals not only elucidate essential system information but also contribute to the reliability of the decision-making process.

10.1 Background & problem formulation

Visualisation of data is ubiquitous and has permeated all areas of research and society. Visualisations capitalise on our fundamental capacity for visual processing to offer versatile and intuitive methods of information processing and representation **(Purchase 2014)**. Networks are commonly used to model relationships between individual components in a complex system and the process of visualising networks is part and parcel of encapsulating a systems-level understanding **(Vehlow et al. 2017)**.

Network modelling offers a comprehensive strategy to delineate the complexities underlying big data. It provides a detailed depiction of the system at hand, capturing the intricacies of interactions among various entities. In network graphs, nodes symbolise the constituent elements of the system, while edges signify the interactions between these elements **(Gibson et al. 2013)**. Typically, network visualisation can be achieved using node-linked diagrams visualised within a type of Cartesian framework **(Y. Wang et al. 2015)**. Cartesian framework is the planar space formed within two perpendicular axes between which numerical data is represented. However, in node-linked diagrams representing networks, the widths and size of the links and nodes take preference over the position of the nodes and the distance between them.

When visualising node-linked diagrams within a Cartesian space, complexity frequently emerges with the increase in the number of constituent elements and their interactions in the network models **(Arleo et al. 2017)**. This complexity often inhibits clear communication. Part of the challenge stems from the fact that the edges tend to increase on a different scale than the nodes. To mitigate this challenge, numerous solutions have been proposed with an aim to enhance the aesthetic appeal and information clarity of these visualisations **(B. Bach et al. 2016; Vehlow et al. 2017; Y. Wang et al. 2015)**. Nonetheless, these methods are often labor-intensive, necessitating extensive manual curation. Furthermore, these methods can be challenging to apply for a less experienced user. This not only introduces potential user biases but also poses significant scalability challenges when generating multiple networks that need to be compared against each other. In numerous network analysis approaches, an essential feature is a capacity to distinguish differences in the behavior of intricate systems across various scenarios. This necessitates the visualisation and comparison of multiple networks simultaneously **(Jahagirdar and Saccenti 2020b)**. Consequently, it becomes challenging to scale labour-intensive network visualisation techniques in these types of applications. Some tools have been developed to allow for extemporaneous synchronised visualisation of networks **(Lindfors et al. 2018)**, however, the *anvis* package extends this synchronisation further by assessing the cumulative information when automating node placements and integrating user-interests via user-determined categorisation for system constituent elements (nodes).

Historically, data and information visualisation have been underpinned by two fundamental principles: reduction and optimal distribution of spatial variables **(Manovich 2011)**. The principle of reduction implies the employment of graphical objectives, encompassing elements such as points, lines, curves, and basic geometric shapes, as means to symbolise objects of interest and their inherent relationships. Meanwhile, the principle of optimal distribution of spatial variables refers to the utilisation of spatiotemporal aspects, including the position, size, shape, and movement

of the said graphical objects, to embody select characteristics and properties of the represented objects **(Manovich 2011)**. We have developed an Automated Network VISualisation package named *anvis* with the objective to automate the decision-making process associated with the selection of spatial variables in network diagrams, also referred to as the network topology. This automation would subsequently enable the generation of multiple human-readable co-ordinate visual representations of networks, without the necessity for manual curation.

Network visualisations of intricate systems serve a dual purpose: they not only elucidate and communicate information visually but also serve as an invaluable tool for informed decision-making **(Dyson 2016; Borkin et al. 2013)**. This tenet was particularly highlighted in **Chapter 9** *Wisdom of the Informed Crowds*. In this context, our software package, *anvis*, crafts data-driven representations of network models utilising data science and machine learning methodologies (regressions and sigmoid curves). This substantially expedites the production of publication-ready network visualisations, simultaneously reducing user bias and enhancing the reliability of decisions based on these figures. A key design philosophy behind *anvis* was to facilitate seamless access for novice users to high-quality graphics, while simultaneously offering experienced users comprehensive access to the underlying tools and algorithms. This enables users to customise the package according to their specific needs, without necessitating a reverse engineering approach. Consequently, *anvis* can be seamlessly integrated into any user-defined analysis pipeline.

10.2 Application

The software package under the name of *anvis* is programmed in R and strictly adheres to the rules, guidelines, and recommendations outlined by Bioconductor. *anvis* is currently available open source on GitHub under the GNU General Public License v3.0. It can be installed from https://github.com/VanderJag/anvis.git.

10.2.1 Installation

The package can be installed from GitHub following the installation instructions detailed here and on the GitHub page. The package can install all the dependencies from CRAN, however, the user is required to install the following dependencies from Bioconductor themselves: RCy3, BioNet, BiocStyle. Bioconductor can be installed by running the following code in R:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```

The Bioconductor dependencies can then be installed using the following R code:

```
BiocManager::install(c("RCy3", "BioNet", "BiocStyle"))
```

The *anvis* package can then be installed from the GitHub repository using devtools. If required devtools can be installed by running the R code:

```
if (!requireNamespace("devtools", quietly=TRUE))
    install.packages("devtools")
```

The *anvis* package can then be installed by running the following code in R:

```
devtools::install_github("VanderJag/anvis",
    build_vignettes = TRUE)
```

The code will install *anvis*, its dependencies, and also build the vignette. The vignette can be accessed using the following code:

```
vignette("Vignette", package = "anvis")
```

The package is able to visualise networks internally in the R environment using igraph **(Csardi et al. 2006)** and also externally using Cytoscape **(Shannon et al. 2003)**. To utilise the later functionality, Cytoscape needs to be installed from https://cytoscape.org/download.html and kept open in the background. For a seamless experience, we recommend installing the latest Cytoscape version or the version after 3.9.0.

10.2.2 Inputs and outputs

A fundamental objective during the development of the *anvis* package was to ensure its capacity for seamless integration into any user-defined data analysis pipeline. This has meant that the package accommodates a broad assortment of input and output selections. The package can read either a single network object or a list of multiple network objects representing multiple networks. Furthermore, in the resulting visualisations, we have ensured alignment with user needs by adopting user-determined categorisation for system constituent elements (nodes). Therefore, the package is capable of interpreting a vector encapsulating the labels of the groups into which the nodes are organised. Concerning outputs, the package can generate images in a spectrum of common formats, as well as export networks in specific formats that retain the network layout and topological variables. These can be subsequently imported into a network visualisation software of the user's preference for visualisation containing the user's style preferences. Figure 10.1 highlights a comprehensive list of input (highlighted in green) and output (highlighted in light pink) formats.

10.2.3 Description & core functionalities

The *anvis* package is designed to automate the creation of network visualisations, employing data-driven visual styling to emphasise the inherent behaviour of the system. It proves especially advantageous for producing multiple network comparisons,

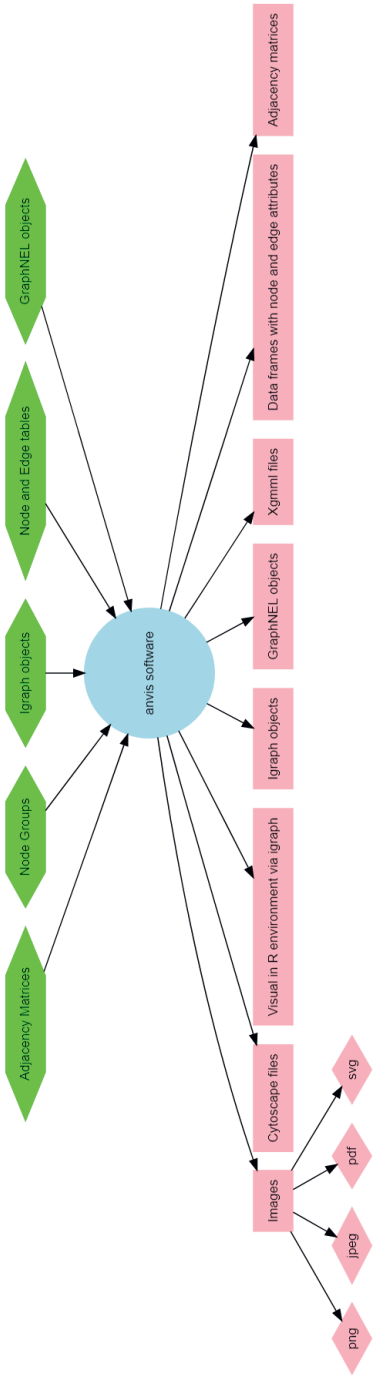


Figure 10.1: Inputs and outputs of the *anvis* software package: The inputs are given in green ant the outputs are highlighted in light pink. The images can be exported in all commonly used image formats. Eventhough the figure only lists 4 highly used formats, the package allows the user to access all image formats supported by base R, igraph and cytoscape.

enabling the study of diverse states or scenarios in the underlying system. The package allows for user-determined node categorisations to align the visualisations with the user's interests. This user-established categorisation impacts both the positioning and colour coding of the nodes within the network.

The majority of the package's primary operations can be conveniently executed via three wrapper functions: "anvis", "addVisAttrs", and "adjToNetwork". The "anvis" function takes a (list of) network object(s) and visualises them or exports them in formats specified in figure 10.1 preserving the network layouts and variables describing the network topology. The function allows for the visualisations to be made using the igraph package in R or cytoscape (an external software). To facilitate easier comparisons across multiple networks, "anvis" can depict several networks as sub-plots within a user-defined grid. The default layout of the networks is a circular layout and the node positions are kept constant across all the visualised networks. The radius of the circular layout can be adjusted based on the empty space kept between each node with the argument cyto_node_space and the edge widths can be scaled using the argument vis_edge_factor. Detailed information can be found in the package documentation **(https://doi.org/10.5281/zenodo.8128358)**.

Operating synergistically with the "anvis" function, "addVisAttrs" is designed to append node and edge attributes to the networks that are to be visualised. This function brings a substantial degree of personalisation to the network visualisation. It enables users to fully customise node and edge size, colour, position, width, and categorisation while offering default options. For instance, by setting the argument "colorblind=TRUE", users can ensure that network visualisations employ a colour palette conducive to colour-blind viewers. This function also allows the implementation of entirely custom colour palettes. In the visualisation, edge widths are modelled on a sigmoid curve between the network's minimum and maximum interaction values. The "width_type" argument provides users with a selection of methods for modelling edge widths. Options include "MI", "cor", "partcor", "ranked", and "percentile", thus enabling a tailored visualisation to suit the user's hypothesis. The software also allows for the assignment of distinct colour palettes to positive and negative edge values. Detailed information is available in the software's documentation, with examples provided in the vignette.

Networks are often depicted as matrices, where row and column names correspond to the nodes, and matrix values signify undirected or directed edges depending on the matrix's symmetry. The "adjToNetwork" function transforms these (weighted) adjacency matrices into network objects, which can then be visualised by the "anvis" function. While constructing the network object, "adjToNetwork" offers numerous arguments for associating relevant attributes with the nodes and edges. These include the "width_type" and "colorblind" arguments discussed previously, which are used to model edge widths and select colour-blind friendly palettes, respectively.

10.3 Usage & examples

Numerous figures in this thesis can serve as good examples as they have been visualised using this software, or an earlier version of the software. These examples include Figures 2.8, 2.9, 4.3, 4.5, 5.3, 5.6, 6.12, 6.13, 7.2, and 8.13. However, we pro-

vide examples that show the usage of package here. Accompanying the software package are two datasets that underpin the examples demonstrated herein. The first dataset, denoted "paleo", comprises matrices where nodes symbolise trace elements gathered from a deep-sea drilling experiment. This data has been analysed to explore environmental shifts during the late Paleocene to early Eocene epochs in Earth's history. An extensive description of the data can be found in **(Thomas et al. 2005)**. The second dataset, labelled "sepsis", features matrices that depict networks with nodes embodying blood plasma proteins measured in hospital patients afflicted by soft tissue infections. An extensive description of the data can be found in **(Medina et al. 2021; Rath et al. 2023)**

In this section, we offer a selection of examples illustrating the package's application. Detailed instructions and the corresponding codes for all functionalities are available in the package vignette. The initial example, depicted in Figure 10.2, involves a directed network visualised using the "paleo" data. For illustrative purposes, the code for this example assumes that the data resides on the user's local storage, thereby mirroring the process a user would undertake when applying the package to their own dataset.

```
#Example 1: Directed paleo network
# set of commands used to generate figure 10.2

#Load the data from your local storage directory.
#Make sure the matrix has column names and row names
#and they are identical.
paleo <- as.matrix(read.table("Local_storage/paleo.csv",
    sep=",", header = TRUE, row.names=1))

#Classify the column names/nodes into groups.
#Here we simply use the labels
#GroupA, GroupB, and GroupC
#Nodes from each group will be positioned together
#and get assigned the same colour.
G <- as.vector(c("GroupA",
    rep("GroupB", 6), "GroupC",
    rep("GroupB", 6),
    rep("GroupA", 2), "GroupC", "GroupA")

#Now use the adjToNetwork function to convert the matrix
#into a graphNEL network object for visualisation
net1 <- adjToNetwork(paleo, directed = TRUE,
    self_loops = FALSE, node_attrs = "all",
    edge_attrs = "all", group_vec = G,
    size_type= "cytoscape", width_type = "percentile")

#Visualise and save the network
vis <- anvis(net1, directed = TRUE,
    save_names = "paleo_network_directed",
    output_type="cytoscape",
    vis_save = TRUE, vis_edge_factor = 1)
```

The second example depicted in Figure 10.3 shows a grid of 20 networks created from the sepsis dataset showing the protein interactions in different cases. Here the interactions are positive (in shades of red) and negative (in shades of blue) in nature. In this example, we load the data available directly from the package.

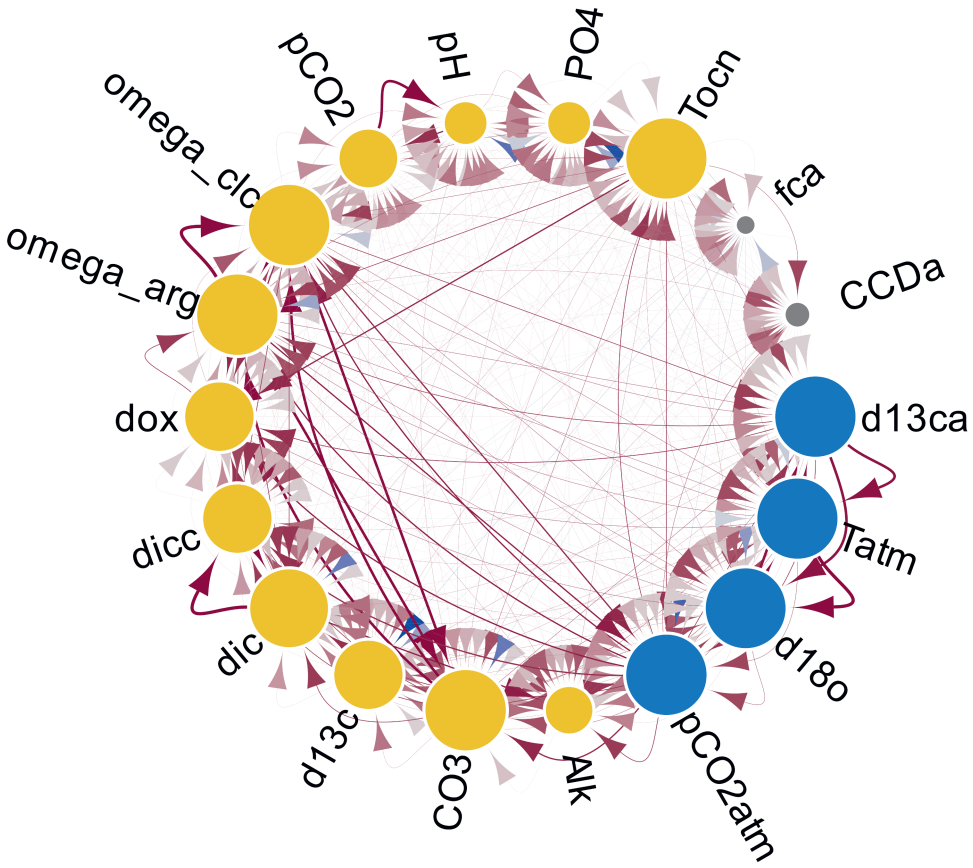


Figure 10.2: A directed network visualised by *anvis* using the paleo dataset. The three different groups are in three different colours and the direction of the edges is given by the direction of the arrows. The edges in shades of red are positive interactions and the ones in the shades of blue are negative interactions.

```
#Example 2: A grid of 20 networks from sepsis data

#Sepsis data is available with the package
sepsis <- c(sepsis[1:12], sepsis[1:8])

#Here once again we make some groups
#and label them GroupA, GroupB, GroupC, and GroupD
#extract column names
proteins <- colnames(sepsis[[1]])
#Group the proteins based on their names
groups <- dplyr::case_when(
    stringr::str_starts(proteins, "IL") ~ "group A",
    stringr::str_starts(proteins, "CCL") ~ "group B",
    stringr::str_starts(proteins, "CXCL") ~ "group C",
    TRUE ~ "group D")

#Convert the list of adjacency matrices into
#a list of network objects, with
#all additional attributes added.
net_list <- adjToNetwork(sepsis,
                node_attrs = "all",
                edge_attrs = "all",
                width_type = "partcor",
                group_vec = groups)

#Give names to each individual network
grid_titles <- c("CaseA", "CaseB", "CaseC", "CaseD",
    "CaseE", "CaseF", "CaseG", "CaseH", "CaseI",
    "CaseJ", "CaseK", "CaseL", "CaseM", "CaseN",
    "CaseO", "CaseP", "CaseQ", "CaseR", "CaseS",
    "CaseT")

#Visualise the networks
#igr_grid = c(5,4) creates a 5 by 4 grid
anvis(net_list, igr_grid = c(5,4),
      igr_par_opts = list(mar=c(2,4,5,4)))

#we adjusted the margins of the networks using igr\_par\_opts,
#so the node labels are shown completely.
```

While the default layout of the package is a circular layout, the design of the package allows the same principles and algorithms to be applied to diverse network layouts. In our third example, showcased in Figure 10.4, the visual aesthetics are applied to a Davidson-Harel layout algorithm, which is based on simulated annealing **(Davidson et al. 1996)**. The fourth example, visible in Figure 10.5, illustrates the visualisation of multiple networks to compare distinct cases using a grid-based algorithm.

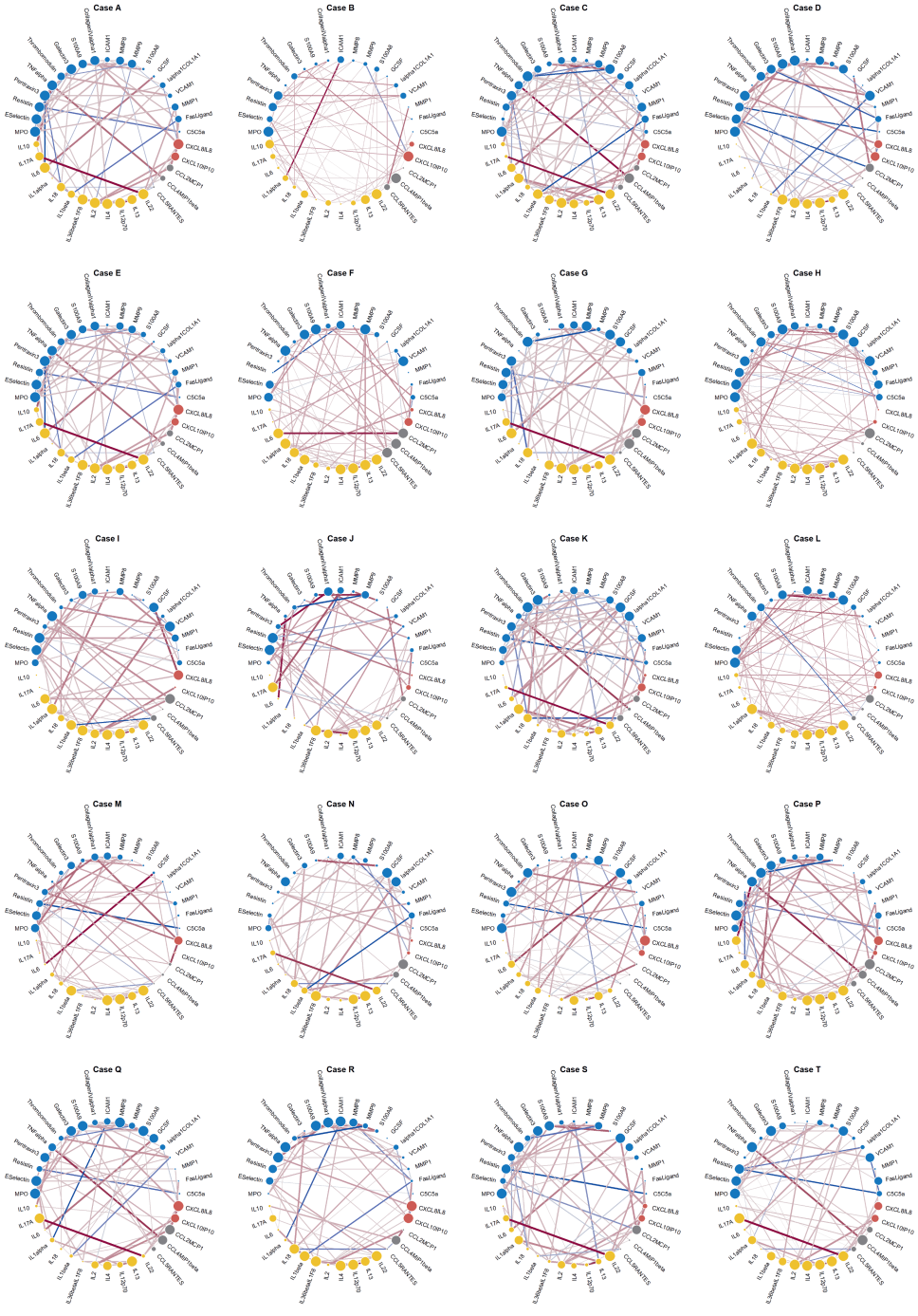

Case A
Case B
Case C
Case D
Case E
Case F
Case G
Case H
Case I
Case J
Case K
Case L
Case M
Case N
Case O
Case P
Case Q
Case R
Case S
Case T

Figure 10.3

Figure 10.3: 20 networks visualised simultaneously with the same amount of programming steps in a ready-to-publish 4 × 5 grid. The groups of the nodes are in distinct colours and the node positions of the proteins stay exactly the same in each of the 20 networks, thus making comparisons intuitive. The size of the node also indicates the importance of that node in the network. Positive and negative interactions are differentiated by shades of red and shades of blue respectively.

```
#Example 3: Davison–Harel layout algorithm

#Since the sepsis data is loaded and
#the matrices are converted to network objects,
#we can use the same networks.

#Visualise the first network object in the list
#using Davison–Harel layout algorithm by passing
#the visual options to igraph through the argument
#igr_plot_opts
anvis(net_list[[1]],
    igr_plot_opts = list(layout = igraph::layout_with_dh,
    vertex.label.cex = 0.6,
    vertex.label.color = "black",
    vertex.label.family = "sans"),
    vis_radial_labs = FALSE)

#Example 4: Multiple networks in grid–layout

#We can create a function to manually alter the
#number of rows and columns in the grid layout
layout_func <- function(graph) {
    igraph::layout_on_grid(graph,
    width = 4,
    height = 9)}

#Visualise the first 4 networks in the sepsis data
#using the grid layout
anvis(net_list[1:4],
    igr_plot_opts = list(layout = layout_func,
    vertex.label.cex = 0.7,
    vertex.label.color = "black",
    vertex.label.family = "sans",
    asp = 1.4),
    igr_par_opts = list(mar=c(0,1,1,1)),
    vis_radial_labs = FALSE)
#We adjusted appearance of the node labels,
#aspect ratio of the node grid,
#and margins of the plots for a more clear image.
```

10.4 Future development

The *anvis* package serves as an effective initial step in the visualisation of extremely large networks, providing a foundation upon which other packages, such as NetBioV **(Tripathi et al. 2014)** or yfiles (Cytoscape), can be subsequently applied. Nevertheless, when tasked with visualising exceptionally large networks, the *anvis* package's capabilities as a stand-alone tool are limited. We have adopted a modular design in the package to facilitate the future integration of models into its pipelines. A prospective endeavour involves developing an XGBoost-based modelling pipeline that embodies principles delineated in **(Kwon et al. 2017; Vehlow et al. 2015)**. This advancement will considerably enhance the *anvis* package's capacity to automate the visualisation of extremely large networks.

10.5 Conclusion

Network visualisation serves as a pivotal step towards achieving a systems-level comprehension. However, the production of high-quality, human-readable visuals demands a significant manual curation effort. In response to this challenge, we have devised an Automated Network VISualisation software package called *anvis*. This package automates the decision-making process linked with the selection of spatial variables that delineate the network's topology. Consequently, *anvis* is capable of generating multiple, data-driven, human-readable visual representations of networks while allowing for a substantial degree of personalisation to the network visualisation.

10.6 Funding

The study received funding from the Wageningen Data Competence Center (WDCC) via the Data Science & Artificial Intelligence fellowship.

10.7 Author contributions

SJ designed the study and acquired the funds; RK designed the software; SS provided data and validation by testing the software; MSD, ES, IVdL VAdP provided resources; SJ wrote the manuscript; All authors provided critical revision and reviewed the manuscript.

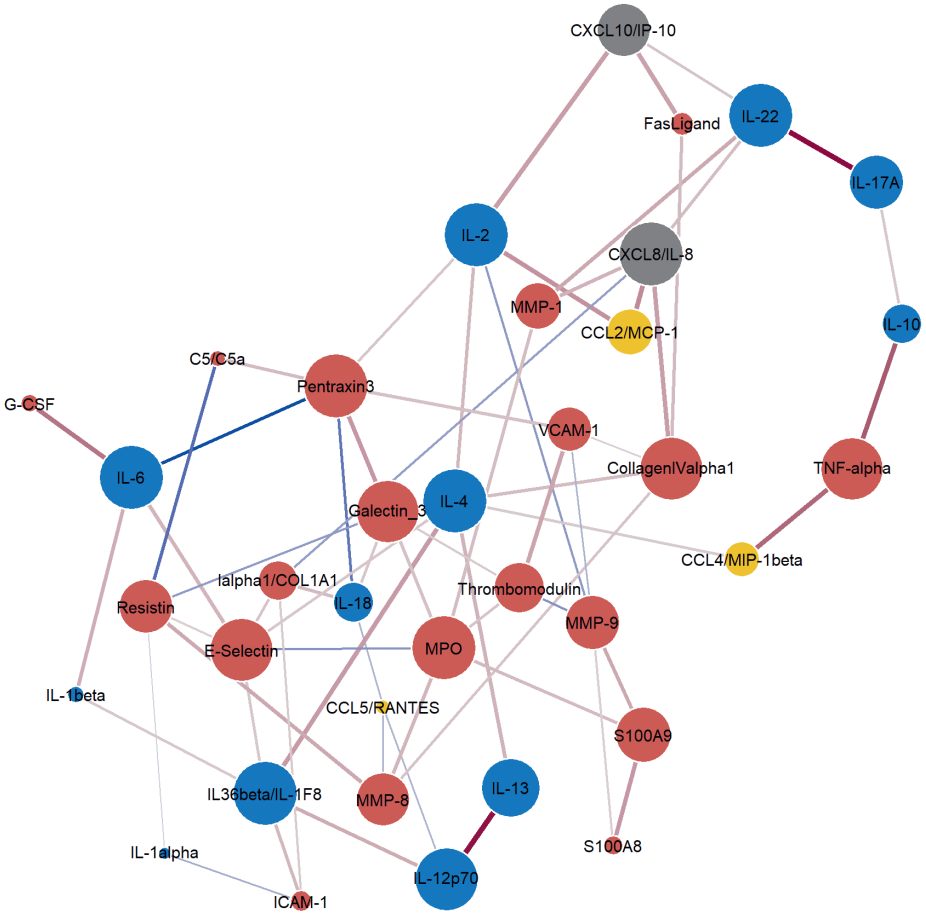


Figure 10.4: Network visualised by *anvis* from the sepsis dataset using the Davison-Harel layout algorithm **(Davidson et al. 1996)**

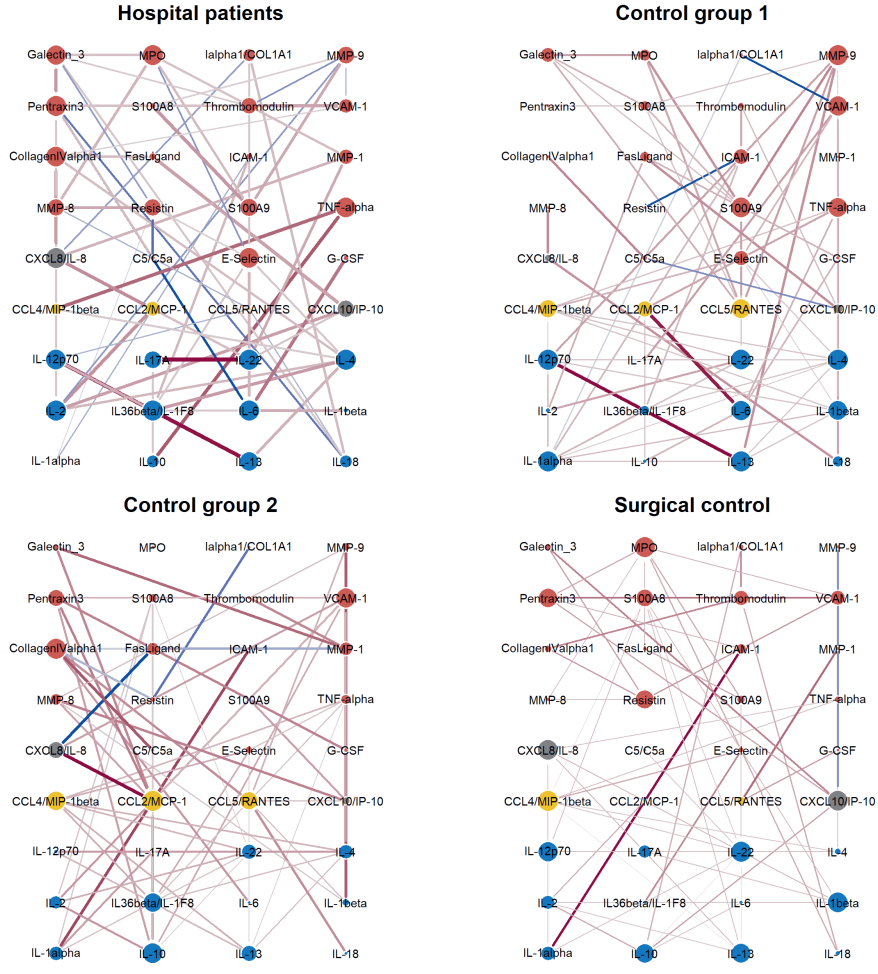


Figure 10.5: Four networks visualised simultaneously from the sepsis dataset in a grid-based layout. Once again the node positions stay the same in all the networks making it easy to differentiate between them.

Chapter 11

Chapter

Sanjeevan Jahagirdar[1]

Turn to page 377 for author affiliations

Discussion: Piecing it Together

The broader aim of this thesis has been to advance the use of computational modelling, data science, and machine learning methods in the biomedical research surrounding NSTI in order to explore and understand the underlying mechanisms while identifying variables with predictive and discriminatory power in distinguishing between NSTI, sub-types and controls. NSTI (Necrotising soft tissue infections) are a predominantly bacterial disease associated with tissue necrosis, sepsis, high mortality rate and severe loss of quality of life **(Hua et al. 2022; D. L. Stevens and Bryant 2017; Suijker et al. 2020)**.

Towards this aim, **Chapter 2** and **Chapter 3** engage in a detailed exploration of the decisions and intricacies that come into play when network-based methods are employed in the realm of systems medicine. Subsequent chapters **Chapter 4**, **Chapter 5**, **Chapter 6**, **Chapter 7**, and **Chapter 8** focus their attention towards systematically processing the heterogeneous data collected from NSTI patients, approached in a coherent and linear trajectory. Lastly, **Chapter 9** and **Chapter 10** delve into the inherent issues that surface within this multi-/inter-/trans-disciplinary research and offer tools devised to ameliorate gaps in this field.

In **Chapter 1**, I delve into the intricate process of diagnosing NSTIs. This complexity originates from the heterogeneous nature of patients and their symptoms, further exacerbated by the sub-optimal definitions currently used in NSTI characterisations and the need for a speedy diagnosis. Yet, the struggle to precisely define diseases is not an issue exclusive to NSTIs. Indeed, formulating a comprehensive definition of what encompasses a "disease" presents a significant challenge in its own right. Here, I seek to initiate a discourse on the fundamental queries raised when classifying diseases. While Hesslow's objections raised in the concept of disease eliminativism are acknowledged **(Hesslow 1993)**, the author of this thesis firmly believes in the necessity for researchers to diligently scrutinise, contemplate, and reflect on the very implicit assumptions underlying any scientific inquiry.

11.1 The disease dilemma

Diseases encompass a broad spectrum of physical and psychological abnormalities stemming from various sources, including viral, fungal, and bacterial infections, nutrient deficiencies, and genetic or physiological aberrations at the level of genes, proteins, or organ systems. The inherent diversity among diseases is conspicuous; they can exhibit significant disparities in symptoms, debilitating effects, duration, and underlying biological aetiology. Consequently, a pertinent question arises: Is the

quest to define "disease" worthy of engaging in a rigorous scientific or philosophical discourse? or a malformed nonsensical endeavour, better answered with absurdist humour, akin to the whimsical numerical response of "42" **(D. Adams 2010)**?

11.1.1 Disease eliminativism

Hesslow suggests that the concept of disease is superfluous and what ultimately matters is, the desire for beneficial medical intervention and whether the medical institution possesses the power to help **(Hesslow 1993)**. Apart from questioning, what constitutes a medical institution, this publication inevitably raises an additional, ostensibly apparent query: Can an individual competently self-diagnose? It is worth noting that societal consensus would definitely tend to diverge from a potential affirmative individual perspective on this matter. In social species, diseases often tend to not just affect individuals (biological abnormalities) but also societies (epidemiology, public health, economy, research spending) **(Arah 2009)**. The history of such medical institutions is not without its blemishes either. In the absence of precise definitions, medical institutions have often prescribed draconian treatments for individuals showcasing divergence from the general population, especially in cases of neurodivergence. In this context, the author of this thesis dares to imagine what potential pros and cons could be discussed when prescribing lobotomy to a person diagnosed as having "disorganised thinking" **(Rees 1949; Terrier et al. 2019)**?

11.1.2 Conceptualisation of disease

Social animals have developed various strategies to strike a balance between the collective advantages in disease prevention and the potential ramifications of losing active members of the group **(Rueppell et al. 2010)**. Such behavioural adaptations are not only observed in hominins **(Trinkaus et al. 2017; Tilley 2015; Oxenham et al. 2009)** but also in social insects **(Bos, Lefèvre, et al. 2012; Bos, L. Sundström, et al. 2015)**. The ongoing discourse on the conceptualisation of disease often pivots around the distinct considerations at either end of this equilibrium. Most conceptions tend to gravitate towards the philosophical paradigms of either naturalism (or objectivism) or normativism (or constructivism) **(Powell et al. 2019)**. The principal distinction between these two schools of thought lies in the emphasis placed on the influence of value-laden judgements.

11.1.3 Naturalism

The naturalist prospective is well captured by Boorse's bio-statistical theory where he defines a disease (later re-coined as pathological function **(Boorse 2014)**) as a state of statistically species-subnormal biological part-function relative to a reference class **(Boorse 1977)**. This definition invites reflection on the feasibility of a disease afflicting an entire population; a recurring theme in science fiction novels, films, games, and other entertainment media. It also prompts examination of real-world examples, such as the three-toed sloth's toxic diet **(Pauli et al. 2014)** despite its distinctive ancestry **(Presslee et al. 2019)**, or the prevalence of sickle cell anaemia genes in regions afflicted by malaria **(Aidoo et al. 2002)**. One could also debate the existence of such ideal biological functions assigned to systems in the light of evolutionary exhibits

the likes of a 28 metre-long recurrent laryngeal nerve in *Supersaurus* that regulated the animal's swallowing mechanism **(Wedel 2011)**. However, as poised by Kingma, more significant scrutiny is deserved regarding the selection of the reference class and the potential introduction of value judgements during such a selection **(Kingma 2010)**. Furthermore, Boorse's conceptual framework appears to categorise diversity in human physiology, such as homosexuality or neurodivergence, within the scope of disease. While it is appropriate to categorise a subset of deviations, such as developmental disorders, as diseases, it is also worthwhile to contemplate the notion that we (*Homo sapiens*) exist only because of neotenous sponges **(Romer 1967)**.

11.1.4 Normativism

Normativist characterisations grounded in holistic and quality of life perspectives often manage to tackle this issue of diversity **(Goosens 1980)**. However, quality of life perspectives are a function of expectations and people with different expectations tend to report different quality of life for the same clinical conditions **(A. J. Carr et al. 2001)**. Furthermore, their tendency to render disease classification as contingent on spatiotemporal and cultural contexts, coupled with their inability to biologically/biomedically condemn morally problematic cases (historical examples include draptomania **(Bynum 2000)** or female sexual activity outside of marriage **(Pasko 2010; Alexander et al. 2006)**) as abuses of the disease concept is less than ideal **(Powell et al. 2019)**.

11.1.5 The power of narratives

While this section introduces attempted classifications and their impact on the boundary cases within the philosophical landscape, infectious diseases and by extension NSTIs have been straightforwardly classified based on their causal aetiologies, thanks largely to the advancements in germ theory. Be that as it may the case, an important aspect warrants introduction, one that affects most characterisations of disease: the influence and power of narratives and stories. Green et al. have demonstrated that stories exert a persuasive influence on public narratives, resulting in reduced identification of flaws and the expression of beliefs consistent with the story's context **(Green et al. 2000)**. Moreover, narratives possessing the power to captivate individuals have been found to enhance the transportation effect, leading to heightened emotional processing and a cognitive realignment with the narrative's constructed fictional realm **(Appel et al. 2007)**. This effect has also been shown to increase over time with the aptly termed sleeper effect **(Appel et al. 2007)**. In recent decades, narrative medicine has become an accepted part of medical treatment that tends to give voice to patients suffering from the disease and their prospective **(Charon 2008; Cartledge et al. 2020)**.

However, patients do not hold a monopoly on such narratives. These stories and narratives encompass various emotional frameworks, including the perspectives of the storyteller and listener, the temporal progression, and plot structures that may strategically portray individuals or objects in the light of victims, rescuers, or persecutors, effectively enhancing the story's persuasive appeal. Terminology such as "war on coronavirus" **(Draghi 2020)** or "war on cancer" **(Sporn 1996)** is frequently

employed to rally societal and resource investment towards specific objectives. This binary, adversarial narrative (us vs them), despite straying from empirical evidence, is readily digestible by both the scientific community and the broader public **(Clark et al. 2019)**. I would propose that this style of storytelling influenced the formation of the self vs. non-self model of the human immune system and the introduction of descriptors like "good bacteria" and "bad bacteria" when reality differed from the model. Of course, in recent years, a far more plausible model is used with improved capacity to explain aspects of the human immune response **(Matzinger 2002)**. While the aforementioned case could be considered a testament to the value of ongoing scientific discourse, research, and progression, I would personally like to reflect on the potential impact of narratives such as "homosexuality is unnatural", "individuals with ADHD are lazy", and "obesity is a personal failure of diet control" have had on individuals despite evidence to the contrary **(Vasey 2002; Flanigan 2021; Kyrou et al. 2009)**. My personal opinion is that a society failing to commemorate those who diverge from the general population risks bypassing the significant long-term benefits that emerge from the expansion of ideas, divergent thought processes, and extension of knowledge barriers, which may be the outcomes of the varied and different experiences these individuals bring to the collective.

Even Hesslow, in his paper, presents an analogy that draws a parallel between a perceived acceleration issue in a car and a biomedical problem in a patient. He criticises the philosopher for engaging in a debate about the disparity in the definition of mechanical faults employed by the car owner and the mechanic, suggesting instead to solely focus on discussing the pros and cons of modifying the engine valves **(Hesslow 1993)**. However, it is important to question whether something that holds true for a car, ergo must it hold true for a human being? Additionally, it is worth noting that this narrative can be modified in various ways to support virtually any conceivable concept.

In **Chapter 1** of this thesis, specifically in the section *Necrotising Soft Tissue Infections*, the distressing tale of Frankie, a one-year-old boy diagnosed with NSTI, serves as an impactful introduction. This narrative, reproduced from **(Cartledge et al. 2020)**, aims to engage readers and instil a keen interest in further exploration of NSTI and the importance of this thesis. It's crucial, however, to recognise the existence of a bias here; as evident from table 5.1, the median patient age for both monomicrobial and polymicrobial infections stands at 57 and 55 respectively, a finding corroborated by additional studies **(M. B. Madsen, Skrede, et al. 2019)**. Thus, the more common scenario might involve a 60-year-old patient with multiple comorbidities, which could arguably draw less attention than the story of a young child facing a life-threatening misdiagnosed infection. Despite the lack of malice behind this bias, readers should remain cognisant of its influence and the emotional impact carried by this narrative and other scientific narratives presented throughout this thesis.

11.2 Exploring the solution of stratification

The work done in this thesis underscores the marked heterogeneity observed in the characteristics, symptoms, and aetiology of NSTI patients, as previously emphasised in different chapters of this thesis. This diversity has engendered a range of defini-

tions and terminologies applied to NSTIs, thereby inhibiting the establishment of a wide consensus.

11.2.1 Inherent heterogeneity in NSTI

As elaborated in the preceding section *The disease dilemma*, disease manifestation can be a multifaceted interplay of aetiology, habitat, environment, and host responses. It is important to acknowledge the limited symptomatology that may represent a significantly broader spectrum of potential perturbations.

Certain diseases offer a straightforward path to diagnosis, however, diseases like NSTIs present a unique challenge due to their heterogeneous aetiologies, disease signatures, and effects and are possibly classified into NSTI only with some unifying feature inherent to all patients diagnosed with NSTIs.

11.2.2 Stratification of patients

Addressing the heterogeneity among patients could be accomplished through stratification, grouping individuals based on similarities in patient characteristics and disease signatures. This presents a pragmatic approach to reducing the complexity inherent in patient cohorts.

Any proposed system of patient stratification needs to leverage the cumulative knowledge concerning microbial aetiology, disease sub-types, location of infections, geographical distribution, patient characteristics, and disease signatures. The anticipated outcome would be a marked progression towards achieving the objectives of personalised medicine, precision medicine, and tailored immunotherapies. Additionally, such an approach holds the potential for optimising the utilisation of medical resources and hospital infrastructure, consequently enhancing the cost-effectiveness of healthcare systems. Ultimately, such a stratification scheme has the potential to set a foundation for a more targeted, efficient, and cost-effective NSTI healthcare landscape.

Chapter 2 evaluates single sample networks, specifically their potential to discern patient-level network perturbations and stratify patients using blood metabolomics data. In **Chapter 4**, we endeavour to delineate differential host-pathogen interactions in patients stratified by both microbial aetiology and infection location. In **Chapter 5**, the stratification of patients is pursued through hierarchical clustering applied to patient-specific perturbed interactions, particularly those involving host-pathogen interplay.

Table 4.3 highlights a substantial conditional dependence of microbial aetiology on the site of patient biopsy used for data acquisition. In line with this, table 4.3 underscores that all top-ranking genes exhibiting conditional dependencies with microbial aetiologies are derived from *S. pyogenes*, despite the data containing genes from various species and the expectation of polymicrobial infections caused by diverse bacterial species.

Figures 5.3 and 5.4 illustrate that gene interactions in *S. pyogenes* are suggestive of immune evasion strategies, such as complement inhibition, while interactions with other species indicate their recognition and identification by the human immune

system. This includes genes associated with the recruitment of natural killer cells or activation of the complement cascade.

In table 5.4, we identify the *S. pyogenes* gene P0C0H1 (HasA), which encodes for hyaluronic synthase, a recognised virulence factor **(DeAngelis et al. 1994)**, interacting with the human gene encoding Collagen VI. This interaction is noted in stratified patient groups 2, 3, 5, and 6 during monomicrobial infections. Furthermore, Collagen III interacts with the *S. pyogenes* gene A2RGM6 (Putative two-component response regulator), a protein known to regulate emm, another notable virulence factor **(Ribardo et al. 2004)**. Collagen V is observed to interact with another *S. pyogenes* gene, F5U6Q2, in patients of group 3. In all polymicrobial sub-groups, interactions are observed between Collagen XV and *P. micra* genes.

11.2.3 Differentiating between type I and type II NSTI

Playing the role of an idea man, I propose, based on our data and findings, that the distinction between Type I and Type II Narcotising Soft Tissue Infections is not exclusively dictated by microbial aetiology. It is plausible that the patient's individual characteristics, response, and the location of infection play a substantial role in the observed differences between these two NSTI types.

Further, I conjecture that the immune evasion strategies utilised by *S. pyogenes*, coupled with patient-specific characteristics and comorbidities, as well as the internal environment in the context of the infection site, may inadvertently cultivate an ideal milieu. This environment could potentially encourage the proliferation of bacteria typically deemed harmless under normal conditions.

Skrede et al. have also advocated for classifying aetiology into primary and secondary microbes, based on their capability to instigate an infection irrespective of comorbidities, as discussed in **Chapter 1 (Skrede et al. 2020)**. However, as noted by **(Hua et al. 2022; D. L. Stevens and Bryant 2017)**, this understanding does not concretely impact the diagnosis and initial treatment of NSTI patients. Nevertheless, grasping these underlying systems proves crucial for long-term management and development.

11.2.4 The need for experimental validation

While this argument holds biological plausibility, it should serve primarily as a springboard for formulating hypotheses that require further investigation. This principle applies not only to the currently proposed hypothesis but also to numerous data-driven hypotheses suggested throughout this thesis in **Chapter 4**, **Chapter 5**, **Chapter 6**, **Chapter 7**, and **Chapter 8**.

Our current comprehension of the biological systems that underpin the human body and the intricate interactions that occur with pathogens falls short of allowing us to make credible predictions on biomechanics. The mechanistic metaphors we often use, such as the lock and key mechanism or metabolic pathways, align with our brain's understanding at the human scale of space and time. This often provides us with a false sense of understanding of biological mechanisms often backed up by years of repetitions of the same concepts and words. However, we struggle to inherently grasp mechanisms at extremely large or small temporal and spatial scales

and repetition of a concept does not make it true. This is discussed in more detail in the next section *In the shadows of causation: why correlation isn't enough?*

Historically several biologically plausible mechanisms such as using Intravenous Immunoglobulin (IVIG) **(Shankar-Hari et al. 2012)**, activated protein C **(Bernard et al. 2001)**, and Vitamin C **(A. C. Carr et al. 2015)**, had been suggested for sepsis drug development. Nonetheless, IVIG showed no effect against NSTI in a randomised, blinded, placebo-controlled trial **(M. B. Madsen, Hjortrup, et al. 2017)**. Moreover, Xigris, the highly touted drug based on activated protein C, was withdrawn from the market following a second double-blind, placebo-controlled, multicenter trial that demonstrated a complete lack of beneficial effect, despite numerous severe side effects **(Gaardlund 2006; E. Abraham et al. 2005)**. Finally, the use of vitamin C supplementation is controversial at best and the impacts are inconclusive at worst **(Ahn et al. 2019)**.

This thesis offers a range of biologically plausible, data-driven hypotheses related to NSTI diagnosis, identification, and underlying mechanisms, spanning from immune evasion strategies to aetiology-dependent responses, from diagnostic biomarkers to the pro- and anti-inflammatory behaviour of cytokines. These factors should be considered in any stratification attempts. However, it is crucial to approach these hypotheses with due scepticism, acknowledging the necessity to corroborate them with future experimental evidence.

11.3 In the shadows of causation: why correlation isn't enough?

In 2012, Messerli published a study in the New England Journal of Medicine, delineating a linear relationship between national chocolate consumption and the enhancement of cognitive function as indicated by the proxy measure of per capita Nobel laureates within a country. His analysis suggested that Switzerland's high count of per capita Nobel laureates was attributable to elevated chocolate consumption. Additionally, Messerli proposed that Sweden constituted an outlier in this analysis, potentially due to national bias within the decision-making committees. In the spirit of full disclosure, Messerli stated a personal conflict of interest, revealing his daily consumption of Lindt's dark chocolate varieties **(Messerli 2012)**.

While this publication may carry an air of jest, it simultaneously serves as a serious study. Masserli's work does not attempt to assert a linear relationship between chocolate consumption and enhanced cognitive function. Instead, it emphasises that correlation does not imply causation, an interpretation that some may have overlooked **(Linthwaite et al. 2013; Kayser 2012)**.

11.3.1 Association does not imply causation

Generally, when one speaks of correlation, it denotes Pearson's correlation, a measure of association initially designed to determine the relationship between two variables based on their normalised co-variance **(Pearson 1895)**. However, similar to how certain brand names such as "Xerox", "Google", "Band-Aid", and "Thermos" have undergone a process of genericization in certain socio-cultural contexts, signifying

photocopying, online searching, adhesive bandages, and vacuum flasks, respectively, the term "correlation" has also experienced a similar scientific genericization by using it to denote general associations **(Clankie 2013)**. It is true that the terms "association" and "correlation" evolved concurrently throughout history, but "correlation" has developed a more defined meaning, particularly in connection with the product moment correlation coefficient introduced by Pearson **(Pearson 1895)**. Consequently, the commonly used adage, "correlation does not equate causation" would perhaps be more accurately phrased as "association does not equate causation" despite its reduced rhetorical appeal. The motivation behind this statement stems from observing a certain degree of personal frustration when researchers employ various association measures to suggest causal inferences. This tactic can occasionally circumvent the scrutiny of the peer review process, specifically due to the omission of the term "correlation". It's noteworthy that while journals might have explicit editorial policies governing the use of correlation measures, these may not be uniformly applied to analogous association measures.

11.3.2 Human biases in determining causal explanations

As articulated by Okasha, uncovering the causes of natural phenomena is a central objective in scientific inquiry **(Okasha 2002)**. This provokes fundamental questions about the nature of causality and the methodologies for inferring causal relationships from data. The rationale behind the assertion "correlation does not imply causation" often revolves around the potential role of confounding variables. These variables may induce a perceived statistical relationship between two distinct events, yet the true causal factor may be altogether different. In Masserli's example, a plethora of factors such as per capita income, economic freedom, work-life balance, among others, might offer more comprehensive explanations for the observed correlation. Sloman et al. argue that our cognitive system inherently perceives causation as the primary driver of events **(Sloman et al. 2015)**. As examined in section *The disease dilemma*, narrative structures play a substantial role in the formation of our understanding of causality as well. According to the prevailing psychological framework, known as the story model **(N. Pennington et al. 1986; N. Pennington et al. 1992)**, decision-making processes hinge on the formation of narratives rooted in complex causal knowledge about both physical and mental causation **(Klein 2017)**. These narratives are constructed to conform to individuals' assumptions about the intentional, goal-directed actions of agents, facilitated by the organisation of events into a coherent causal and temporal process. The story model integrates both narrative and dependency-based knowledge, with the latter offering the general causal knowledge necessary to devise specific narratives **(Sloman et al. 2015)**. These narratives string together actual causes leading to a conclusion. By focusing on intentional causality within a narrative framework, the story model effectively encapsulates a key aspect of how individuals conceptualise causality **(Sloman et al. 2015)**. However, narratives only play a part in the causality puzzle.

The foundational considerations of causal constructs, forming an integral part of one's thought processes have their origins in the work of **(Michotte 2017)**, who embarked on an exploration of causality detection within human perception. this line of research was broadened by **(P. A. White 2014)** through the introduction of

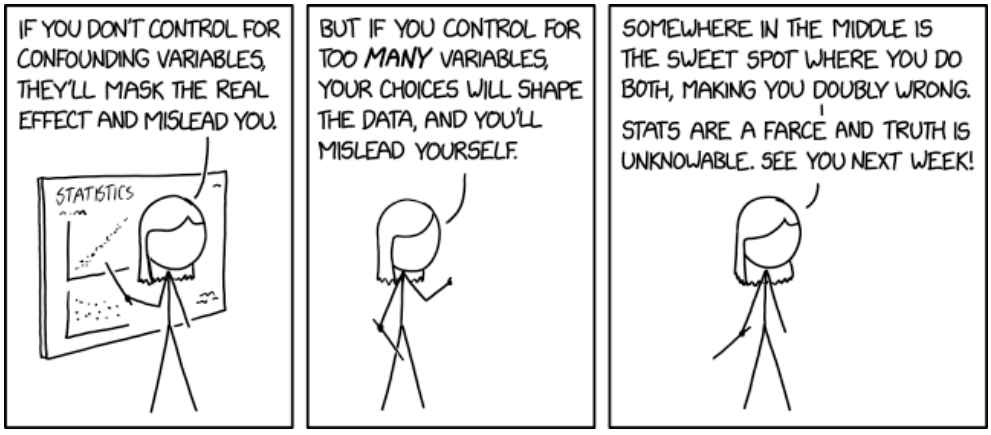


Figure 11.1: A cartoon by xkcd (a webcomic) depicting the fruitlessness of chasing causal inferences. The comic is reproduced unaltered here under the Creative Commons Attribution-NONCommercial 2.5 License. The original cartoon can be found via this permanent link: https://xkcd.com/2560/

a heuristic set designed for the perceptual system to identify causal relationships. Notably, research has demonstrated the capacity of causal relationships to influence, and potentially distort, our perception of events, including perceived time delays and individuals further recognising the sequence of events to align with their anticipated timeline dictated by their causal models, even if the causal beliefs were acquired recently**(Sloman et al. 2015; Buehner et al. 2009; Buehner et al. 2010; Bechlivanidis et al. 2013)**.

11.3.3 Causal relationships or probabilistic dependencies?

Given the inherent biases that influence our understanding of causal inference, it is worth interrogating whether it is appropriate to conceptualise human knowledge as primarily composed of causal relationships between properties and events, or if it might be more accurate to consider it as probabilistic dependencies. The question also arises as to where the responsibility of determining causality rests. Does it fall on the results gleaned from data analysis? or on experimental outcomes? or does it emerge through the cycle of hypothesis-driven systems biology **(Kitano 2002)**? Moreover, could the scientific consensus, cultivated through the collective wisdom of the crowd over time, serve as a determinant? If the consensus approach does hold sway, is there a potential to expedite this process through ensemble-based analysis pipelines, such as **(Setty et al. 2023)**? Further exploration of this notion can be found in the subsequent section *High-level considerations before embarking on an expedition of exploratory modelling*. Furthermore, if the limiting factor for determining causation was just confounding variables, analyses following the logic of partial correlations would be enough to remove spurious associations **(Linn et al. 1969)**. Yet if the aim is to understand mechanistic concepts in biology, these methods are not the most stringent. When studying mechanisms underlying NSTIs, which may include several causes for certain observed symptoms (such as the numerous immune evasion meth-

ods of *S. pyogenes*) and very low sample sizes, making a causal claim is susceptible to the phenomenon of explaining away or discounting **(Pearl 1988; Kelley 1973; M. W. Morris et al. 1995)**. If one of the causes is known to have occurred, and the effect is observed, the probability of all the other causes reduces and it is easy to come to the conclusion that one particular cause caused the effect. This may potentially explain the irreproducibility of LRINEC, a diagnostic method proposed for NSTI detection in 2004. LRINEC and its irreproducibility were introduced in the section *Diagnosis* of **Chapter 1**.

11.3.4 A mechanistic view of biology promotes causal interpretations

If the causal inference is chasing the understanding of mechanistic concepts, it is probably important to also explore what we mean by biological mechanisms. In biological sciences, according to **(Ross 2021)** the concept of "mechanisms" proves instrumental to the elucidation of various biological phenomena across various domains, from gene regulation to visual processing. Mechanistic explanations often invoke an analogy with mechanical systems or machines, thus emphasising the role of constituent parts, their spatial organisation, and their causal interactions in producing a system's behaviour **(Ross 2021)**. This once again underscores the point I raised in the section *Exploring the solution of stratification* that we as humans inherently struggle to understand mechanisms at extremely large or small temporal and spatial scales and metaphoric descriptions have their limits and can give a false sense of understanding. Mechanisms are characterised by three key traits: a constitutive makeup, implying that higher-level behaviours can be decomposed into lower-level causal components; extensive causal detail, wherein systems are described in comprehensive causal terms rather than abstract notions; and an emphasis on the "force," "action," or "motion" within causal relations, offering a richer understanding of how a mechanism operates **(Ross 2021)**. Sedgewick et al. built an MGM using an ensemble of undirected graph learning based on a log-likelihood model and graph search methods based on conditional independence and a stability selection procedure to make a claim of achieving causal inference **(Sedgewick et al. 2019)**. Others have used Bayesian network representations and graphical models to claim causal inference **(Koller et al. 2009; Runge et al. 2019)**. Many machine learning models have also been utilised to claim causal inference.

11.3.5 Probabilistic dependencies were calculated stringently in this thesis

In this thesis, I do not make any claims related to causality or causal inference. As discussed in this section, I believe this to be an unrealistic claim that is affected by several biases discussed above. However, I apply methods that are just as statistically stringent or even more stringent than the methods making a causal claim. In **Chapter 4** I use an ensemble of MGM models, with one focusing on the joint probability distribution between variables with a lasso penalty and the other focusing on the strict conditional independence using a random forest algorithm with stability selection. In **Chapter 5** I use an estimate of partial correlation and a robust resampling approach

to calculate a probabilistic measurement of edge likeliness. Furthermore, the probabilistic measure is corrected for multiple testing using the Benjamini & Hochberg method and the new probabilistic measure was used as a confidence level to reject any partial correlation with a probability below 0.95. In **Chapter 6** and **Chapter 7**, a large ensemble of univariate, multivariate, machine learning and network modelling approaches were utilised with the highest stringency to assess the data and the negative results under the high stringency were reported as negative results. In **Chapter 8** the ratios and parameters were analysed with the XGBoost model 1000 times and the mean, 95% CI and the best values were all reported before deriving conclusions from the results. Even though people find causal reasonings more natural than other forms of reasonings for problems with the same complexity **(Cummins 1995; Cummins et al. 1991)**, I personally believe that we should focus more on probabilistic stringency and ethical consideration and let a concept of causality develop over the consensus of several studies in due time.

11.4 An unexpected data-driven luminary: Thrombomodulin

To the best of my knowledge, prior to the publications, manuscripts, and chapters presented in this thesis, there have been no published reports exploring Thrombomodulin in NSTI patients. The mean concentration of Thrombomodulin in NSTI patients from the INFECT cohort is 11332 ± 2985 pg/ml as reported in table 6.1, far higher than any of the control groups. Notably, Thrombomodulin has been extensively studied in relation to sepsis. Its pivotal role in coagulation and inflammation processes in sepsis has been established **(Levi and Van Der Poll 2012)**. Furthermore, recombinant Thrombomodulin demonstrated efficacy in reducing both 28-day and in-hospital mortality in a subset of severe sepsis patients presenting with severe coagulopathy, elevated fibrinogen/fibrin-degradation-products, D-dimer levels, severe organ dysfunction, and high mortality risk **(Kudo et al. 2021)**. However, the literature presents varying observations, with some studies noting the usefulness of Thrombomodulin in sepsis due to decreasing levels **(Iba, Hagiwara, et al. 2017; Kudo et al. 2021; Yamakawa, Ogura, et al. 2013)** while others report increasing levels **(Kinasewitz et al. 2004; Iba, Yagi, et al. 1995; J.-J. Lin et al. 2017; Mihajlovic et al. 2015)**, likely reflecting the heterogeneity of sepsis patients and the pathology itself.

11.4.1 Thrombomodulin is a good discriminator between NSTI and non-NSTI

Chapter 6 identifies Thrombomodulin as a unique biomarker with the potential to discriminate between NSTI and non-NSTI cases. Moreover, it demonstrates the ability of Thrombomodulin to differentiate at an early disease stage (Figure 6.6). It is important to interpret these findings cautiously, given that the classification of disease stages is subject to bias unavoidable in patient chart notes. The iterative ROC analysis reveals a very high AUC of 0.95 (95% CI [0.89-1]), indicating a strong ability of Thrombomodulin to differentiate NSTI from non-NSTI. The analysis proposes a

threshold of 7566.85 pg/ml for this classification, with higher concentrations suggesting NSTI corresponding to a sensitivity and specificity of 0.92 and 0.89, respectively (Table 6.5). Similarly, the Random Forest models for NSTI and non-NSTI differentiation (based on protein concentrations alone or in combination with relevant clinical variables) identify Thrombomodulin as a key predictor, with significant decreases in both accuracy (217.09 & 111.07 with clinical variables) and Gini index (6.36 & 2.54 with clinical variables) (Table 6.5, p-value=0.01). Thrombomodulin shows substantial discriminatory power between NSTI and non-NSTI cases, as evidenced by q-scores of less than 0.005 in both Kriskall Wallis and Mann-Whitney U tests (Figure 6.5). A similar observation is made in the validation cohort of **Chapter 6** (Table 6.9 and Figure 6.11) where Thrombomodulin is able to differentiate between sepsis patients and NSTI. However, in this case, Thrombomodulin values in sepsis were elevated albeit not to the levels observed in NSTI. In **Chapter 8**, we see the ratio CCL-4/MIP-1β $\iff$ Thrombomodulin as a strong discriminator, such that a ratio below 0.10 is associated with NSTI. Furthermore, α_2IL-4 $\Rightarrow$ Thrombomodulin and α_4IL-4 $\Rightarrow$ Thrombomodulin are also identified as parameters with strong discriminatory ability between NSTI and non-NSTI. Notably, IL-4 has been shown to counteract the downregulation of Thrombomodulin triggered by TNFα and IL-1 **(Kapiotis et al. 1991)**. As seen in Figure 6.12 of **Chapter 6**, IL-4 exhibits a strong positive association with CCL-4/MIP-1β.

In **Chapter 7**, we observe an AUC value of 0.908 from the ROC analysis when differentiating streptococcal NSTI from cellulitis, with a Benjamini-Hochberg corrected p-value of less than 0.005 from the Mann-Whitney U test (Table 7.3). Although the Random Forest model indicates a high mean decrease in the Gini index, the corresponding p-value is 0.06, slightly above the standard arbitrary threshold accepted for significance (0.05). It should be noted that the samples in this study contain a low number of severe cellulitis cases, which could potentially affect the significance of the observed differences. I personally don't believe it would be a stretch to think that the difference would be more significant if there were more severe cellulitis cases in the study.

11.4.2 Thrombomodulin is associated with bacterial toxins

As documented in **Chapter 4**, the concentration of Thrombomodulin appears to be conditionally dependent on markers of sepsis severity, including creatinine, lactate, and platelet levels. Additionally, Thrombomodulin is found to be associated with bacterial genes that respond to environmental stressors and regulate host body temperature and fever (HrcA (Q5XAD4), DnaK (U9WVJ8)). Thrombomodulin is also linked to optimal *S. pyogenes* biofilm formation through the potent toxin $\beta - NAD^+$-glycohydrolase and is associated with *S. pyogenes* proteins functioning as Zn scavengers, thereby weakening the human immune response through Zn deprivation.

In **Chapter 5**, we document several bacterial genes interacting with different collagen molecules. These interactions can trigger the coagulation pathway, promoting platelet activation and aggregation. Figure 6.12 reveals variable associations between Thrombomodulin and Collagen-IVα1 in NSTI patients and surgical controls, with the former showing no association, while the latter exhibits a strong positive association.

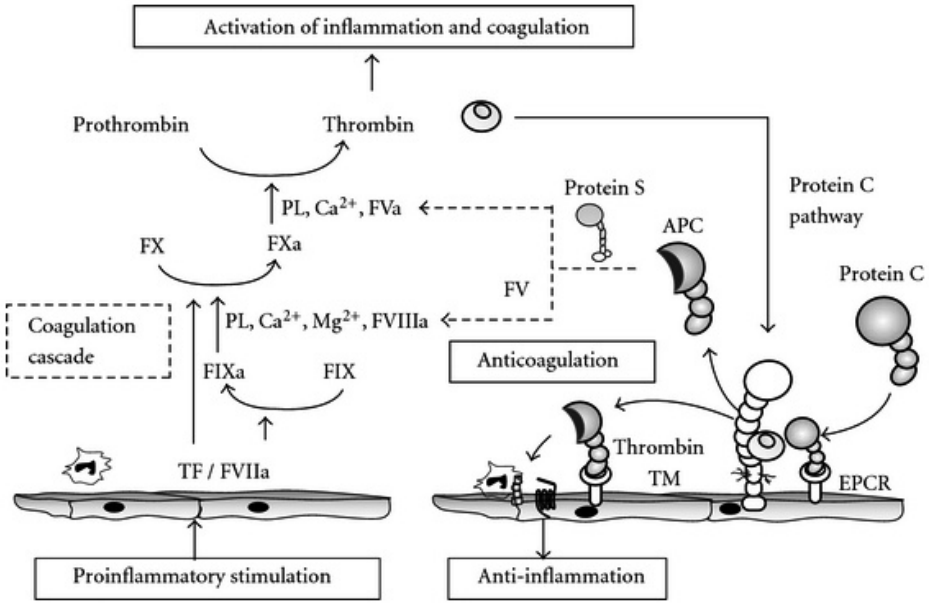


Figure 11.2: An illustration of the reciprocal interplay between inflammation and coagulation **(Okamoto et al. 2012)**. Produced unmodified from **(Okamoto et al. 2012)** under the Creative Commons license 3.0. Proinflammatory stimulation triggers TF expression, initiating a coagulation cascade that eventually leads to thrombin generation **(Okamoto et al. 2012)**. Thrombin further boosts coagulation and inflammation, forming a positive feedback loop. Conversely, the thrombin-TM complex, through Protein C activation, inhibits coagulation and inflammation, establishing a negative feedback loop. Central to this process are the roles of thrombin, TM, and the Protein C-EPCR complex.**(Okamoto et al. 2012)**

11.4.3 Anti-inflammatory effect of Thrombomodulin

Thrombomodulin exhibits dual mechanisms of anti-inflammatory activities, hinging on both activated protein C (APC)-dependent and independent processes **(Okamoto et al. 2012)**. In the APC-dependent paradigm, Thrombomodulin accentuates the activation of protein C, resulting in the suppression of inflammatory cytokine production by inhibiting the nuclear translocation of NF-$\kappa\beta$ components **(B. White et al. 2000; Joyce et al. 2002)**. This modulation leads to enhanced endothelial barrier function and restricted neutrophil migration into inflamed regions **(Elphick et al. 2009; Okamoto et al. 2012)**. In **Chapter 4** Thrombomodulin was found to be associated with Dnase C (A2RDE7), a *S. pyogenes* protein that plays a role in the destruction of neutrophil extracellular traps (NETs) (introduced in **Chapter 1**, section *The human immune response*) and dampen the human immune response by limiting phagocytosis and reducing TLR-9 signalling. Furthermore, an association network built on NSTI patients in Figure 6.12 showed a strong positive association between Thrombomodulin and VCAM-1. VCAM-1 is involved in neutrophil rolling and **Chapter 7** Figure 7.1 clustered Thrombomodulin together with VCAM-1. On the other hand,

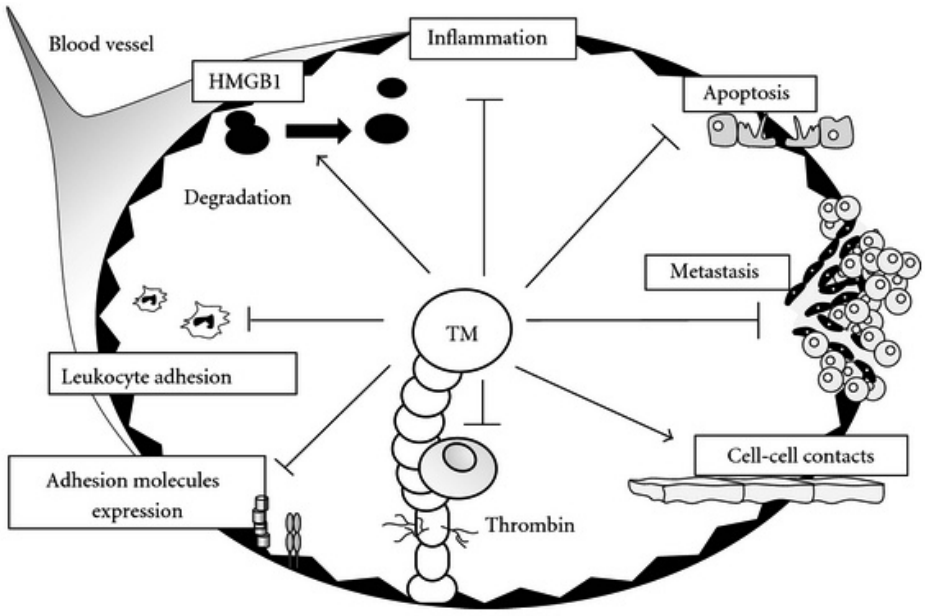


Figure 11.3: The multifaceted roles of Thrombomodulin (TM) in maintaining vascular homeostasis **(Okamoto et al. 2012)**. Produced unmodified from **(Okamoto et al. 2012)** under the Creative Commons license 3.0. TM contributes to the inhibition of inflammation, apoptosis, tumor metastasis, thrombin function, adhesion molecule expression, and leukocyte adhesion. Additionally, the role of TM in neutralizing and degrading the proinflammatory molecule, high-mobility group box 1 (HMGB1) is also shown.**(Okamoto et al. 2012)**

the APC-independent pathway features Thrombomodulin interaction with thrombin activatable fibrinolysis inhibitor (TAFI), which impedes fibrinolysis and mitigates inflammation **(Okamoto et al. 2012)**. Further, Thrombomodulin's lectin-like domain operates independently of APC **(Conway 2012; C. Esmon 2005)**, curbing thrombin's pro-inflammatory effects **(Coughlin 2005)** and disrupting the HMGB1-RAGE signalling pathway **(Abeyama et al. 2005; Okamoto et al. 2012)**. Collectively, these mechanisms underscore Thrombomodulin's crucial role in modulating inflammatory responses and maintaining cellular homeostasis.

11.4.4 Thrombomodulin is an anticoagulant

Thrombomodulin acts as a potent anticoagulant, suppressing the activities of thrombin and accelerating the production of activated protein C (APC) (Figure 11.2) **(C. T. Esmon et al. 1982; N. Esmon et al. 1983; C. T. Esmon 1993; Castellino 1995)**. Thrombomodulin's function extends to inhibiting most of thrombin's procoagulant effects and promoting thrombin's inactivation through the involvement of antithrombin and the protein C inhibitor **(Okamoto et al. 2012; Yang, Manithody, et al. 2003; Preissner et al. 1987)**. This mechanism (via Thrombomodulin) represents a pivotal shift

in thrombin's substrate specificity towards protein C **(T. E. Adams et al. 2006)**. Notably, inflammation can cause a decrease in Thrombomodulin expression (induced by TNFα **(Moore et al. 1989)**), thereby intensifying blood coagulation **(Okamoto et al. 2012; Moore et al. 1989)**. Similar experimental observation has been reported in sepsis patients, where TNFα and IL-1 have been shown to downregulate Thrombomodulin **(Levi and Poll 2008)**. In **Chapter 6** Figure 6.4, as expected, we see the concentration of Thrombomodulin reduce in Non-NSTI patients (who experience inflammation) compared to surgical controls. A similar effect is observed with Cellulitis patients in **Chapter 7** table 7.5 compared to surgical controls. However, we see the exact opposite in NSTI patients with Thrombomodulin concentrations significantly higher. Furthermore, in **Chapter 6** Figure 6.6 we see that Thrombomodulin concentration increases from the early to late stage of NSTI. Consequently, an association network built on NSTI patients in Figure 6.12 showed a conspicuous absence of association between Thrombomodulin and TNFα. Furthermore, none of the reported Thrombomodulin downregulating cytokines (TNFα, IL-1α, and IL-1β) show up as significant discriminators in any of the models using concentrations, ratios, or parameters in **Chapter 8**. Studies suggest that Thrombomodulin plays a protective role in mitigating intravascular thrombus formation, further highlighting its significant role in coagulation pathways **(Okamoto et al. 2012; Isermann et al. 2001)**.

11.4.5 Thrombomodulin and septic shock

The interlinking mechanisms of systemic inflammation and coagulation cascades constitute the central pathology underlying septic shock **(Zeerleder et al. 2005; C. T. Esmon 2005)**. Inflammation, facilitated by the synthesis of thrombin through tissue factor (TF), stimulates blood coagulation **(Mackman 2009)**. This effect is upregulated on monocytes, macrophages, and endothelial cells **(Mackman 2009)**. IL-6 have been shown to induce the expression of Thrombomodulin **(Okamoto et al. 2012)**. This process subsequently leads to fibrin formation and platelet aggregation, resulting in the coagulation of blood **(Okamoto et al. 2012)**. Concurrently, inflammation impedes critical anticoagulant pathways, regulated by antithrombin, protein C, and tissue factor pathway inhibitor **(Okamoto et al. 2012)**. In **Chapter 4** Thrombomodulin is associated with *S. pyogenes* M protein (JZM209), a surface protein involved in antiphagocytic activity. Interestingly, the M protein (JZM209) has been shown to influence the coagulation cascade and increase pro-coagulant activity mediated by the upregulation of tissue factor on monocytes. Furthermore, the coagulation function evoked by the M protein may not be limited to the site of the infection and result in a systemic activation **(Paahlman et al. 2007)**. Interestingly, the coagulation process and anticoagulation pathways also have the capacity to modify the inflammatory response by inhibiting fibrinolysis and inactivating C3a and C5a (proteins part of the human complement system) **(Okamoto et al. 2012; Bajzar et al. 1996; Campbell et al. 2002)**, thereby creating a complex web of interactions that regulate both coagulation and inflammation in septic shock (Figure 11.3) **(Okamoto et al. 2012)**. In **Chapter 4** Thrombomodulin was associated with human genes that highlighted the GO term "platelet-derived growth factor alpha-receptor activity, complement component C4b receptor activity". Furthermore Thrombomodulin was found to be associated with the gene scpA (H8F5G0) that encodes for the immune evasion protein

C5-peptidase. Along with ScpA (H8F5GO), Thrombomodulin was associated with HIT (A2RCZ2) and HIT (J7M9C9), all three of which are involved in reducing the effects of the human complement system.

11.4.6 *S. pyogenes* interaction with Thrombomodulin: part of an immune evasion strategy?

The general anticipation of reduced Thrombomodulin concentrations during inflammation is supported by the lower Thrombomodulin levels observed in non-NSTI, cellulitis, and intermittently (not always) in sepsis patients. These levels are observed to be lower than those of surgical controls. Contrary to this expectation, however, NSTI patients consistently exhibit markedly elevated Thrombomodulin levels.

Despite Thrombomodulin's known anti-inflammatory role, an intense inflammatory response, or hyper-inflammation, is a common feature in NSTI patients. Intriguingly, no discernable difference in the levels of Thrombomodulin downregulating cytokines is observed between NSTI and non-NSTI patients. When considering these findings in conjunction with the relationships between Thrombomodulin and *S. pyogenes* gene expression outlined in **Chapter 4**, it is possible to link Thrombomodulin's discriminatory ability with a myriad of *S. pyogenes* activities designed to attenuate the human immune response.

Considering the full scope of these results, I propose that the increased concentration of Thrombomodulin in NSTI patients may be indicative of an immune evasion strategy implemented by *S. pyogenes*. Furthermore, a deeper understanding of the mechanisms surrounding Thrombomodulin in NSTI could unlock significant advancements in various aspects of NSTI research.

11.5 High-level considerations before embarking on an expedition of exploratory modelling

Introduced in **Chapter 1**, machine learning constitutes a field dedicated to the automation of predictive model creation. Evaluating the predictive capabilities of such models is relatively direct: it involves using an independent test dataset, previously unseen by the model, to evaluate the model's predictive prowess. However, the interpretability of these machine learning models often poses challenges, a conundrum attributable to the nature of the automated modelling process itself. Given the automation, the models must cater to a myriad of features, distributions, and properties inherent in the data being modelled, thereby leading to intricate mathematical representations rendering the model parameters less transparent. This feature also confers a distinct "plug-and-play" characteristic on these predictive models.

There is an acknowledgement of the heterogeneity and diversity in different modelling approaches pursued in the sciences. They are often also recognised as autonomous instruments of inquiry, more than just sole approximations of an underlying hypothesis or a mere aggregation of observations. Traditionally criteria for model evaluations typically include their predictive, explanatory, and representational abilities in relation to specific observations, hypotheses or measurements. However, Gelfert argues the necessity for acknowledging scientific exploration as

an equally vital function of scientific models **(Gelfert 2018)**. Exploratory models have had a positive effect in many different fields operating with complicated systems and irreducible uncertainties **(Kwakkel et al. 2013)** Ofcourse one may not assume that all models should inherently be exploratory but one should not ignore the often-overlooked role of models in exploratory science. In **Chapter 8**, we take an exploratory approach to investigate if different types of pair-wise relationships between pro- and anti-inflammatory cytokines hold any information on the underlying mechanisms of NSTI. Here our aim was not to mimic the precise cytokine dynamics observed in the human body, but instead to investigate if parameters representing both model characteristics and real world concentration data can reveal more information in regards to discriminating NSTI, patient outcomes and microbial aetiology and to interpret the said information in accordance with current literature surrounding NSTI. Such exploratory models, especially computational exploration can have following use cases that are important in general scientific inquiry

- Starting points for further experimental inquiry.
- Providing proof(s) of principle.
- Providing potential avenues for explanation of certain behaviours.
- A tool for reassessing the suitability of the targets in a study or hypothesis.

Gelfert strengthens the case for the use of exploratory models by providing the examples of Turing's biological pattern formation model and Maxwell's molecular vortex model. Both these models were proven to be significantly more useful in due time **(Gelfert 2018)**.

A key challenge within the purview of this thesis lies in facilitating an exploratory discourse to generate data-driven hypotheses that can be pursued further. This exploratory methodology necessitates enhanced interpretability. Moreover, it provides ample latitude regarding the assumptions made during their deployment. This approach pivots away from replicating precise dynamics, instead fostering a more playful exploration of models that align with the collected data.

The development of these exploratory methodologies necessitates an in-depth comprehension and a collaborative endeavour with experts from the application domain. The assumptions underpinning the model must satisfy the feasibility benchmarks inherent in the field, and any subsequent interpretations must be contextualised within the prevailing research corpus and the influence of pre-established assumptions. If not, we run the risk of potentially instigating a concept I would like to call "helicopter-data-science".

The term "helicopter research" or "parachute research" typically refers to the scenario where resource-rich individuals/groups from locations with access to resources or "haves" travel to resource-deprived regions or "have-nots" to gather materials for subsequent processing, analysis, and publication **(Haelewaters et al. 2021)**. While this phenomenon is commonly recognised in the context of developed and developing nations, its scope is far from confined to this setting. Data Scientists, with their ability to access vast arrays of datasets across various disciplines, can rapidly execute their algorithmic pipelines and publish results, often without a comprehensive understanding of the domain. Nonetheless, such a practice may not always contribute

positively to the respective domain, particularly in the absence of specialist input from that field.

Exploratory models, by design, confront inherently "wicked" problems where adequate data collection may be hindered due to ethical, economic, or physical constraints **(Lawrence et al. 2022)**. Even when data are obtainable, these issues may lack a definitive formulation, involve complex interpretations, and the derived solutions may not conform to a binary true-false paradigm. Additionally, researchers often harbor conscious and unconscious preferences towards certain rationales or explanatory models during the assumption process inherent to exploratory modeling. These preferences can encompass empirical accuracy, scope of the explanation, consistency, simplicity, plausibility, precision, quantitative formalism, or explanations that provide guidance for future research **(Brewer et al. 1998)**. Moreover, researchers may rationalise these preferences using established principles. For instance, a preference for simplicity might be justified by invoking Occam's Razor **(Domingos 1999)**. Yet, it is worth noting that Occam's Razor recommends against unnecessary multiplication of entities and does not promote the simpler solution outright. Moreover, defining simplicity is a subjective task, and a universally satisfactory definition is elusive **(Domingos 1999)**. Equally, an overemphasis on accuracy might overlook methods with lower bias or higher computational speed. Therefore, within the context of exploratory models, it is pivotal to embrace diverse methodologies and ensemble modelling techniques and consider the cumulative results.

Another concern that may arise with these exploratory strategies pertains to researchers persistently testing various methods until a significant result is obtained, which is then promptly published. This practice often leads to false positives. To mitigate this issue, it may be beneficial to employ an ensemble-based approach, which incorporates multiple modelling strategies. I advocate for an approach that assesses a cumulative view of all results, including those that are negative, thereby ensuring more robust and accurate conclusions.

In **Chapter 2**, we investigate the properties of single sample networks within an ensemble, which comprises (i) three types of simulated data, (ii) metabolic profiles derived from a dynamic metabolic model simulation, (iii) 22 publicly available metabolic datasets, and (iv) blood metabolite data from NSTI patients. Subsequently, in **Chapter 3**, we evaluate the efficacy of association measures using an ensemble of (i) 23 publicly accessible metabolomics datasets along with 7 datasets from varied domains, (ii) simulated data embodying known correlation structures, and (iii) data generated through a dynamic metabolic model. In **Chapter 4**, we apply a stringent statistical approach employing an ensemble of two MGM models. The first model calculates conditional dependence based on joint probability distribution, incorporating a lasso penalty, while the second determines conditional independence utilising a random forest algorithm with stability selection. Finally, in **Chapter 6** and **Chapter 7**, we adopt an ensemble approach featuring a diverse array of univariate, multivariate, machine learning, and network modelling techniques, taking into account their cumulative outcomes.

11.6 It's time to open Pandora's ethical box

Chapter 1 introduces the pragmatic necessity for distinguishing between NSTI and non-NSTI, from the viewpoint of a medical surgeon. In **Chapter 6**, we elucidate the significance of Thrombomodulin as a key discriminatory biomarker, as determined by the feature importance of a random forest model aiming to differentiate between NSTI and non-NSTI. In **Chapter 8**, models developed using XGBoost demonstrated that the ratio CCL-4/MIP-1β $\Longleftrightarrow$ Thrombomodulin and parameters expressing relationships between Thrombomodulin and other proteins displayed distinctive capabilities in distinguishing NSTI from non-NSTI. **Chapter 4** revealed a conditional dependency of Thrombomodulin on several known virulence factors of *S. pyogenes*. On a similar vein, the discriminatory capacity of CXCL-10/IP-10 between Type 1 and Type 2 NSTI was consistently observed across **Chapter 4**, **Chapter 5**, **Chapter 6**, **Chapter 7**, and **Chapter 8**.

11.6.1 My role in the medical research supply chain

It is paramount here to acknowledge that the aforementioned results from the research conducted in this thesis advances our understanding of NSTI in many dimensions. The results could be taken into considerations for diagnostic and prognostic purposes as well as when building decision support systems. I, however, have never met or interacted with the patients themselves, nor are most of the patients aware of the analysis that I am personally running on intimate data collected from them even if they have consented to the collection and use of their data for research purposes. I just happen to be one node on an increasingly long supply chain.

11.6.2 The promise of Artificial Intelligence methods

AI and Machine Learning methods are continuously being touted for their potential to bolster diagnostic capabilities **(Arieno et al. 2019; De Fauw et al. 2018; Kunapuli et al. 2018)**, drug discovery **(Alvarez-Machancoses et al. 2019; Fleming 2018)**, epidemiology **(Hay et al. 2013)**, personalised medicine **(Barton et al. 2019; Cowie et al. 2018; Dudley et al. 2014)**, and operational efficiencies **(H. Lu et al. 2019; A. Nelson et al. 2019)** in global healthcare systems **(Kube et al. 2019; Morley, Machado, et al. 2020)**. The integration of AI and machine learning in tandem with clinicians is promised to augment medical decision-making abilities with data-driven insights **(Bartoletti 2019)**.

11.6.3 Why ethics matters?

However, piggybacking on this promise are substantial ethical implications that need to be addressed at various stages, from algorithm development and research to implementation, be it in translational research or the creation of decision support systems **(Morley, Machado, et al. 2020)**. A failure to incorporate interdisciplinary perspectives from fields such as medicine, economics, computer science, social science, law, and policy-making could lead to potentially discriminatory and inaccurate outcomes

(Morley, Machado, et al. 2020; Morley and Floridi 2021) thanks to the lack of transparency, accountability and bias-related issues **(Mittelstadt 2019)**. This thesis employs AI and machine learning methodologies for translational research concerning NSTI. Nevertheless, it is worth extrapolating the ethical considerations and deliberations addressed here to the broader healthcare industry, as ethical challenges in NSTI are invariably likely to be very similar.

11.6.4 Data-driven algorithms need to be applied prudently

Data-driven algorithms are frequently lauded for their perceived objectivity, robustness, and adherence to available evidence, often seen as superior to human capabilities in generating diagnostic, prognostic, and treatment recommendations **(Kalmady et al. 2019)**. The axiom "data doesn't lie" has widely infiltrated both scientific and societal paradigms. However, Gillespie et al. critique this presumption as a carefully crafted myth **(T. Gillespie et al. 2014)**. This belief can be not only misleading but can also misconstrue the effectiveness of machine learning algorithms. Indeed, an algorithm's prowess at pattern recognition does not inherently assure meaningful contributions **(Floridi 2014)**. Moreover, these methodologies are often susceptible to overfitting, lack of reproducibility, and limited translatability across different contexts **(Holzinger et al. 2019; Vollmer et al. 2018)**.

Algorithms excel at executing well-defined, precise procedures. However, as explored in section *The disease dilemma*, the definitions of "disease" and "healthy" can be profoundly value-laden. The incorporation of these algorithms into decision-making processes could potentially exacerbate the issues arising from such value-laden definitions of "health," especially when propelled by influential individuals or powerful corporations **(McLaughlin 2016)**. Concurrently, these systems can rapidly magnify standard errors, with the potential to affect not just one patient but hundreds or even thousands in quick succession. This amplification could compound risks associated with misdiagnoses, whether from flawed devices or uncritical acceptance of automated recommendations **(Ruckenstein et al. 2017; Challen et al. 2019)**.

Furthermore, an excessive dependency on these decision-making algorithms may profoundly disrupt the patient-healthcare provider relationship, potentially leading to impersonalisation of care and an overemphasis on quantifiable symptoms. Healthcare providers consider not just measurable symptoms, but also contexts and various observational symptoms. A system heavily reliant on data-driven algorithms may overemphasise quantifiable symptoms and undermine potential alternative diagnostic and treatment approaches **(Juengst et al. 2016; Rosenfeld et al. 2021)**. Medical decisions underscore the necessity for robust information exchange and trust between patients and healthcare providers. However, the "black box" nature of these analytical tools and algorithms can impede this shared decision-making process, potentially excluding the healthcare provider or, more worryingly, undermining a patient's ability and confidence to decline treatment, ultimately infringing upon individual autonomy **(Racine et al. 2019; Sterckx et al. 2016; De Fauw et al. 2018; Ploug et al. 2020)**. Such developments could lead to unintended societal consequences, such as an increasing number of individuals resorting to alternative, non-evidence-based medicine strategies **(Astin 1998)**.

11.6.5 Societal relevance

In **Chapter 1** and **Chapter 11**, specifically sections *The disease dilemma* and *In the shadows of causation: why correlation isn't enough?*, the profound influence of narratives, whether equitable or not, on our understanding of systems and our definitions are emphasised. Such narratives are often anchored in fundamental societal constructs that have evolved with our species. The integration of machine learning and AI systems can potentially foster adversarial narratives, unjustly blaming individuals for not adhering to advice, even when acting in their self-interest **(Morley, Machado, et al. 2020)**. Misplaced blame on individuals, as opposed to systemic shortcomings such as data inaccuracies, algorithmic bias, or unfair treatment among varying demographic groups, would not be unprecedented **(Rhue 2018)**. This could foster the emergence of adversarial narratives that unjustly portray certain groups as morally irresponsible **(Morley, Machado, et al. 2020)**. Individuals may also be encouraged to share personal data for better care, which may be treated opaquely down the long supply chain with no clear accountability and liability for misdiagnosis **(Morley, Machado, et al. 2020)**.

Data plays a crucial role in medical research and the development of hybrid algorithm-based systems, but there is ambiguity concerning the volume and type of data required **(Powles et al. 2017; King et al. 2018; Powles et al. 2018)**. Experts in machine learning often assert that, once data is integrated into their models, only the parameters are shared, leaving the original data unrecoverable. While this claim is not entirely without merit, the advent of quantum computing calls into question its future validity and the potential vulnerability of previous models **(Y. Zhou et al. 2020)**. It is also worth scrutinising whether such technological advancements could expose data transformed via stringent, established processes. Moreover, as explored by Morley et al., the expediency of AI-dependent systems may amplify the impact of flawed assumptions and poor-quality evidence, leading to resource misallocation and possibly initiating a self-reinforcing cycle that could deteriorate public healthcare provisions **(Morley, Machado, et al. 2020)**.

Recent advances in AI and machine learning methods, especially with the development of recent sequence transduction models **(Vaswani et al. 2017)**, hold a great deal of promise for the entire infrastructure surrounding healthcare, from research to applications. However, it is my personal opinion that the ethical qualms can not be ignored and can not be just outsourced to the ethicist. Ethical considerations need to be taken seriously even at the algorithmic developmental stages, and while acknowledging the ethical considerations may not change the optimisation or training methodologies, the biases and limitations need to be properly translated and conveyed along the supply chain.

11.7 Strengths and limitations of the research presented in this thesis

In **Chapter 2**, we evaluated two methodologies developed to discern individual perturbations contributing to the network at large. Our findings indicated that these methods fell short in facilitating generalisations and that the ssPCC approach dis-

played inadequate statistical power. A significant limitation was the inability to distinguish between true perturbations and noise. Potential future studies could address this by contrasting these methodologies against various noise addition procedures. However, the distinction between noise and signal often remains a subjective matter in any scientific discourse. Nevertheless, we found these tools beneficial in exploratory contexts.

Chapter 4 and **Chapter 5** face challenges related to relatively limited sample size, patient heterogeneity in terms of comorbidities, infection timelines, treatments, biopsy sampling differences, tissue types, infection depth, and absence of longitudinal samples, many of these issues are inescapable. **Chapter 6** and **Chapter 7** introduce additional limitations, such as the plasma collection being confined to the study hospitals and the small count of severe cellulitis cases in control groups. **Chapter 8** incorporates further constraints due to the lack of dynamic data on plasma analytes, attributable to mechanistic and ethical considerations.

Regarding computational methods, limitations in **Chapter 4** originate from ranking mixed data types wherein each level contributes either an additive or linear effect when predicting discrete or continuous variables respectively. **Chapter 8** introduces further constraints due to underlying assumptions, the potential for the parameter estimation to be construed as extreme data transformation, and the pro- and anti-inflammatory behaviours of cytokines not being a forgone conclusion.

On the other hand, this approach is fortified by certain quality factors, such as the systematic collection of biopsies by dedicated clinical teams using standardised SOPs, consistent study designs across all participating clinical centres, and standardised plasma collection procedures. Notably, this research involved the largest NSTI cohort to date, a relatively homogeneous patient population with respect to sub-types, and the most extensive GAS and SD patient cohort so far, complemented by comparable control cohorts. The stringent application of statistical methods and ensemble-based modelling approaches enhances the confidence put in the study outcomes. Constant interactions and effective communication in multi-/inter-/trans-disciplinary environments, culminating in the insights presented in **Chapter 9** and **Chapter 10**, further amplify the reliability of the results and future prospects.

11.8 Concluding remarks

Necrotising soft tissue infections (NSTI) are a group of infectious diseases predominantly caused by bacterial species (like *S. pyogenes*, *S. dysgalactiae*, *S. aureus*, and others) that are associated with severe tissue necrosis, sepsis, a high mortality rate, and a significant loss in the patient's quality of life. This thesis sits at the precipice of a multidisciplinary intersection encompassing computational systems biology/medicine, medical research and (micro-)biology with the aim of advancing the use of computational modelling, machine learning and data science methods in the biomedical research surrounding NSTI to explore the underlying mechanisms and identify variables with predictive and diagnostic value in distinguishing NSTI sub-types and controls. Towards this aim, in **Chapter 2** (*Personalising Metabolomics? A Closer Look at Single Sample Network Inference*) we explore the potential use of single sample networks concluding its benefits as an exploratory tool. In **Chapter 3** (*Correlation or*

Mutual Information? That is the Question!) association measures were reviewed in the context of metabolic networks and we found that correlation based measures perform better at inferring true positives in metabolic networks. In **Chapter 4** (*Eventually it Boils Down to the Clinical Variables*), we try to tackle the issues presented with mixed data types and low sample sizes in NSTI by building MGMs (mixed graphical models) with a particular focus on conditional dependence with variables measured in the clinic. We find that the bacteria sense several different environmental clues, based on which change the expression of important virulence factors. We further find several genes related to several known virulence factors such as M protein and LacD1 conditionally dependent on variables measured in the clinic. These genes were further analysed for their interactions with human genes in **Chapter 5** (*Inside the Inferno: An Inspection of the Host-Pathogen Interactions*). Here, we find aetiology-dependent responses highlighting potential modes of entry and immune evasion strategies employed by *S. pyogenes* and interactions with genes associated with signalling proteins called cytokines. Multiple univariate, multivariate, machine learning and network modelling methods were used in ensemble to analyse the concentration levels of these cytokines in **Chapter 6** (*The Race Against Time: Discriminatory Plasma Biomarkers*) and **Chapter 7** (*The Immune System Responds: Systemic Immune Activation Profiles*). Here we identify predictive biomarkers for NSTI clinical phenotypes with potential value for diagnostic, prognostic, and therapeutic approaches. Particularly of note is the identification of Thrombomodulin as a strong differential biomarker between NSTI and non-NSTI, a very important and pragmatic classification as explained in **Chapter 1**. Pair-wise relationships between these cytokines in the form of ratios and model parameters is explored in **Chapter 8** (*Deadly Dance: Understanding the Interplay of Pro- and Anti-Inflammatory Cytokines*). The pro- and anti-inflammatory behaviour of cytokines is taken into account. Several data-driven hypotheses on potential differing mechanisms underlying NSTI is proposed based on the differences between NSTI and controls, patient outcomes, and microbial aetiology. **Chapter 9** (*Wisdom of the Informed Crowds*) draws upon the collective insights gained from the work done in preceding chapters and introduces a problem-solving model tailored for multi-/inter-.trans-disciplinary research. A key component of the model proposed in **Chapter 9** is the utilisation and necessity of human-interpretable, data-driven visualisations, extending beyond mere communication to fortify reliable decision-making processes. Based on this tenet, I was awarded an internal DS/AI fellowship/grant enabling the research carried out in **Chapter 10** (*Automated NETwork VISualisation with anvis*). In **Chapter 10** we develop a software package (anvis) that generates multiple human-readable co-ordinate representations of networks that can be employed with any user-defined analysis pipeline.

To conclude, in this thesis, I address the aim of constructing a structured scientific enquiry using computational approaches to understand the biology of NSTI by (a) advancing the use of computationally derived modelling in **Chapter 2** and **Chapter 3**, (b) promoting several data-driven hypotheses to explore the underlying biological mechanisms in NSTI in **Chapter 4**, **Chapter 5** and**Chapter 8**, and (c) mining heterogeneous multi-omic big data sets by identifying variables of predictive and discriminatory value in **Chapter 6**, **Chapter 7**, **Chapter 4**, **Chapter 5**, and **Chapter 8**. Furthermore, in this thesis, in **Chapter 9** and **Chapter 10**, I endeavoured to develop tools that would provide value to researchers in some aspects of similar multi-

and inter-disciplinary projects. Furthermore, the identified biomarkers, mechanisms and responses influencing phenotypic observations, and several data-driven hypotheses strengthens the rationale for a personalised medicine/systems medicine-based approach in the clinical management of NSTI and underscores the inherent heterogeneity (both from the context of patient and data) of the disease. Finally, identifying these mechanisms helps us take a crucial step in the direction of expanding our diagnostic, prognostic, and therapeutic armamentarium.

Summary

Necrotising soft tissue infections (NSTI) are a group of bacterial infections characterised by widespread tissue destruction in any layer of the soft-tissue compartment ranging from superficial skin layers to the deep musculature. These infections often culminate in high rates of amputation, mortality, sepsis, and a significant decline in quality of life, thereby earning the colloquial designation of "flesh-eating disease". NSTIs are often classified into sub-types based on their microbial aetology. Despite these classifications, there is substantial heterogeneity in reported microbial aetiologies and a lack of international consensus on the definitions and pathogenicity. This thesis aims to construct a structured scientific enquiry using computational approaches to understand and explore the underlying biological mechanisms of NSTI. Towards this aim, (a) I advance the use of computationally derived modelling, data science, and machine learning approaches in the biomedical research surrounding NSTI, (b) use the said approaches to promote data-driven hypotheses exploring the underlying biological mechanisms in NSTI, and (c) employ those computational approaches to mine heterogeneous, multi-omic big data sets with the objective of identifying variables that exhibit predictive and discriminatory value in distinguishing between different NSTI sub-types, aetiologies, and patient outcomes.

Chapter 1 (*Introduction: Why this matters?*) presents an introduction to NSTI and a pragmatic classification of non-NSTI, alongside an overview of systems medicine methodologies and the current landscape of computational techniques within this domain, including network modelling. It also acknowledges the loss of individual-specific information in network models. In **Chapter 2** (*Personalising Metabolomics? A Closer Look at Single Sample Network Inference*), I thoroughly examine two methodologies, ssPCC and LIONESS, developed to restore such lost information, and demonstrate their utility for data exploration using metabolomics data obtained from NSTI patients. Despite the ssPCC method's foundation on Pearson correlation, the separation of the procedure from the inference method (correlation) itself is demonstrated. **Chapter 3** (*Correlation or Mutual Information? That is the Question!*) assesses the effectiveness of correlation and mutual information (as inference methods) in identifying differentially associated metabolites, ultimately concluding that the latter offers no significant advantage over the former in this specific task.

In **Chapter 4** (*Eventually it Boils Down to the Clinical Variables*), I address the challenges associated with mixed data types and limited sample sizes frequently encountered in NSTI data sets, aiming to analyse the relationships between variables observed in clinical and ICU settings, gene transcriptomic measures from humans and bacteria, and protein measurements in plasma. Emphasis is placed on associations with clinical variables, given their accessibility and potential to facilitate prompt diagnosis and treatment. This chapter unveils the complex interactions occurring between *S. pyogenes* and neutrophils, along with the association of Thrombomodulin with markers indicative of severe sepsis and virulence factors recognised for their

interactions with the human immune system and blood constituents. **Chapter 4** further elucidates that *S. pyogenes* modulates the expression of critical virulence factors in response to environmental cues, tying mortality to the regulation of these factors (SpeB, Streptolysin S, and Sic), governed by CcpA and biofilm expression. In **Chapter 5** (*Inside the Inferno: An Inspection of the Host-Pathogen Interactions*), I undertake a comprehensive analysis of these bacterial genes by constructing an interactome using dual RNA-seq gene transcriptomic profiles procured from NSTI patient biopsies. I identify several *S. pyogenes* virulence factors (hyaluronan synthase, Sic1, Isp, SagF, SagG, ScfAB-operon, Fba and genes upstream and downstream of EndoS) interacting with the human stress and immune response. **Chapter 5** reveals aetiology-dependent responses and hypothesises potential entry modes and immune evasion strategies of *S. pyogenes*. The host genes connected with these factors are primarily characterised by their cellular response to cytokines.

An exhaustive analysis of these cytokines is conducted in **Chapter 6** (*The Race Against Time: Discriminatory Plasma Biomarkers*) and **Chapter 7** (*The Immune System Responds: Systemic Immune Activation Profiles*), with **Chapter 6** dedicated to differentiating between NSTI and non-NSTI, while **Chapter 7** concentrates on distinguishing NSTI and Cellulitis. Through an integrated approach, combining statistics, machine learning, and network modelling, discriminatory plasma biomarkers were identified with a potential value for diagnostic, prognostic, and therapeutic approaches in NSTI. Notably, Thrombomodulin is identified as a unique biomarker for distinguishing NSTI from non-NSTI, while differences in the levels of IL-1β, TNFα, and CXCL-8/IL-8 are associated with the differentiation of NSTI and Cellulitis. NSTI severity was associated with a distinctive inflammatory profile, and additional biomarkers were pinpointed for differentiating microbial aetiology (IL-2, IL-10, IL-22, CXCL-10/IP-10, Fas-Ligand, MMP-9) and septic shock occurrence (G-CSF, S100A8, IL-6). **Chapter 8** (*Deadly Dance: Understanding the Interplay of Pro- and Anti-Inflammatory Cytokines*) delves into the information derived from pairwise relationships between these cytokines, plasma analytes, and similar proteins. **Chapter 8** utilises ratios and model parameters as proxies for relationships encapsulating the dual nature of cytokines as pro-inflammatory and anti-inflammatory agents. **Chapter 8** reveals several data-driven hypotheses on potential mechanisms underpinning NSTI. Additionally, the ratio CCL-4/MIP-1β $\iff$ Thrombomodulin is highlighted for it's ability to differentiate between NSTI and non-NSTI.

Chapter 9 (*Wisdom of the Informed Crowds*) integrates both the collective knowledge gained from the preceding chapters and addresses the underreporting of early career researchers' perspectives in solving complex, multi-/inter-/trans-disciplinary issues. In this chapter, we introduce a novel problem-solving model specifically tailored for such research endeavours that prioritises technical aspects without over-reliance on heuristic methods commonly used in individual disciplines. A key component of the proposed model emphasises on the use and importance of human-interpretable, data-driven visualisations, not only as a communication tool but also as a critical component of reliable decision-making processes. Based on this tenet, I was awarded a DS/AI fellowship/grant to create anvis, an automated network visualisation software presented in **Chapter 10** (*Automated NETwork VISualisation with anvis*). This software, engineered to be seamlessly integrated into any user-defined pipeline, employs data science and machine learning methods to automatically gener-

ate multiple, human-readable coordinate visual representations of networks, thereby circumventing the need for manual curation.

Chapter 11 (*Discussion: Piecing it Together*) provides a comprehensive synopsis of the foundational rationale guiding each research segment within this thesis. This chapter delves into the philosophical underpinnings pertaining to societal relevance that shape the primary inquiries addressed within this work. It further probes consequential considerations, ethical implications, and engages in a rigorous critique of the derived results. **Chapter 11** culminates with an introspective evaluation of the strengths and limitations of the conducted research, elucidating their respective influences on the findings presented herein. While postulating several data-driven hypotheses, this thesis highlights the heterogeneity in the host-pathogen interactions in NSTI and furthers the rationale for a systems medicine based/personalised approach in the management of NSTI. The identification of pivotal mechanisms from the hypotheses generated in this thesis have the potential to contribute to expanding our diagnostic, prognostic, and therapeutic armamentarium.

References

Abdillahi, S. M., Maaß, T., Kasetty, G., Strömstedt, A. A., Baumgarten, M., Tati, R., Nordin, S. L., Walse, B., Wagener, R., Schmidtchen, A., et al. (2018). "Collagen VI contains multiple host defense peptides with potent in vivo activity". In: *The Journal of Immunology* 201.3, 1007–1020.

Abeyama, K., Stern, D. M., Ito, Y., Kawahara, K.-i., Yoshimoto, Y., Tanaka, M., Uchimura, T., Ida, N., Yamazaki, Y., Yamada, S., et al. (2005). "The N-terminal domain of thrombomodulin sequesters high-mobility group-B1 protein, a novel antiinflammatory mechanism". In: *The Journal of clinical investigation* 115.5, 1267–1274.

Abraham, A., Pieritz, K., Thybusch, K., Rutter, B., Kröger, S., Schweckendiek, J., Stark, R., Windmann, S., and Hermann, C. (2012). "Creativity and the brain: uncovering the neural signature of conceptual expansion". In: *Neuropsychologia* 50.8, 1906–1917.

Abraham, E., Laterre, P.-F., Garg, R., Levy, H., Talwar, D., Trzaskoma, B. L., Francois, B., Guy, J. S., Bruckmann, M., Rea-Neto, A., et al. (2005). "Drotrecogin alfa (activated) for adults with severe sepsis and a low risk of death". In: *New England Journal of Medicine* 353.13, 1332–1341.

Adams, D. (2010). *The Ultimate Hitchhiker's Guide to the Galaxy: Five Novels in One Outrageous Volume*. Del Rey.

Adams, T. E. and Huntington, J. A. (2006). "Thrombin-cofactor interactions: structural insights into regulatory mechanisms". In: *Arteriosclerosis, thrombosis, and vascular biology* 26.8, 1738–1745.

Afzal, M., Saccenti, E., Madsen, M. B., Hansen, M. B., Hyldegaard, O., Skrede, S., Martins dos Santos, V. A., Norrby-Teglund, A., and Svensson, M. (2019). "Integrated Univariate, Multivariate, and Correlation-Based Network Analyses Reveal Metabolite-Specific Effects on Bacterial Growth and Biofilm Formation in Necrotizing Soft Tissue Infections". In: *Journal of Proteome Research* 19.2, 688–698.

Agarap, A. F. (2018). "Deep learning using rectified linear units (relu)". In: *arXiv preprint arXiv:1803.08375*.

Ahmad, R., Kochumon, S., Chandy, B., Shenouda, S., Koshy, M., Hasan, A., Arefanian, H., Al-Mulla, F., and Sindhu, S. (2019). "TNF-α drives the CCL4 expression in human monocytic cells: involvement of the SAPK/JNK and NF-κB signaling pathways". In: *Cell Physiol Biochem* 52.4, 908–21.

Ahn, J. H., Oh, D. K., Huh, J. W., Lim, C.-M., Koh, Y., and Hong, S.-B. (2019). "Vitamin C alone does not improve treatment outcomes in mechanically ventilated patients with severe sepsis or septic shock: a retrospective cohort study". In: *Journal of Thoracic Disease* 11.4, 1562.

Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., Ter Kuile, F. O., Kariuki, S., Nahlen, B. L., Lal, A. A., and Udhayakumar, V. (2002). "Protective effects of the

sickle cell gene against malaria morbidity and mortality". In: *The Lancet* 359.9314, 1311–1312.
Al Alayed, K., Tan, C., and Daneman, N. (2015). "Red flags for necrotizing fasciitis: a case control study". In: *International Journal of Infectious Diseases* 36, 15–20.
Alamiri, F., Chao, Y., Baumgarten, M., Riesbeck, K., and Hakansson, A. P. (2020). "A role of epithelial cells and virulence factors in biofilm formation by Streptococcus pyogenes in vitro". In: *Infection and immunity* 88.10, 10–1128.
Albrecht, E. A., Chinnaiyan, A. M., Varambally, S., Kumar-Sinha, C., Barrette, T. R., Sarma, J. V., and Ward, P. A. (2004). "C5a-induced gene expression in human umbilical vein endothelial cells". In: *The American journal of pathology* 164.3, 849–859.
Alexa, A., Rahnenfuhrer, J., et al. (2010). "topGO: enrichment analysis for gene ontology". In: *R package version* 2.0, 2010.
Alexander, M., Garda, L., Kanade, S., Jejeebhoy, S., and Ganatra, B. (2006). "Romance and sex: pre-marital partnership formation among young women and men, Pune district, India". In: *Reproductive health matters* 14.28, 144–155.
Alonso-Lopez, D., Gutierrez, M. A., Lopes, K. P., Prieto, C., Santamaria, R., and De Las Rivas, J. (2016). "APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks". In: *Nucleic acids research* 44.W1, W529–W535.
AlShebli, B. K., Rahwan, T., and Woon, W. L. (2018). "The preeminence of ethnic diversity in scientific collaboration". In: *Nature communications* 9.1, 5163.
Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H. J., and Zacharias, H. U. (2020). "Gaussian and Mixed Graphical Models as (multi-) omics data analysis tools". In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863.6, 194418.
Altenbuchinger, M., Zacharias, H. U., Solbrig, S., Schäfer, A., Büyüközkan, M., Schultheiß, U. T., Kotsis, F., Köttgen, A., Spang, R., Oefner, P. J., et al. (2019). "A multi-source data integration approach reveals novel associations between metabolites and renal outcomes in the German Chronic Kidney Disease study". In: *Scientific reports* 9.1, 1–13.
Alvarez-Machancoses, O. and Fernandez-Martinez, J. L. (2019). "Using artificial intelligence methods to speed up drug discovery". In: *Expert opinion on drug discovery* 14.8, 769–777.
Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., and Havel, J. (2013). "Artificial neural networks in medical diagnosis". In: *Journal of applied biomedicine* 11.2, 47–58.
An, Q., Rahman, S., Zhou, J., and Kang, J. J. (2023). "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges". In: *Sensors* 23.9, 4178.
Anaya, D. A., McMahon, K., Nathens, A. B., Sullivan, S. R., Foy, H., and Bulger, E. (2005). "Predictors of mortality and limb loss in necrotizing soft tissue infections". In: *Archives of Surgery* 140.2, 151–157.
Andrade, K., Corbin, C., Diver, S., Eitzel, M. V., Williamson, J., Brashares, J., and Fortmann, L. (2014). "Finding your way in the interdisciplinary forest: notes on educating future conservation practitioners". In: *Biodiversity and conservation* 23, 3405–3423.

Andrews, S. et al. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Appel, M. and Richter, T. (2007). "Persuasive effects of fictional narratives increase over time". In: *Media Psychology* 10.1, 113–134.

Apweiler, R., Beissbarth, T., Berthold, M. R., Blüthgen, N., Burmeister, Y., Dammann, O., Deutsch, A., Feuerhake, F., Franke, A., Hasenauer, J., et al. (2018). "Whither systems medicine?" In: *Experimental & molecular medicine* 50.3, e453–e453.

Arad, G., Levy, R., Nasie, I., Hillman, D., Rotfogel, Z., Barash, U., Supper, E., Shpilka, T., Minis, A., and Kaempfer, R. (2011). "Binding of superantigen toxins into the CD28 homodimer interface is essential for induction of cytokine genes that mediate lethal shock". In: *PLoS biology* 9.9, e1001149.

Arah, O. A. (2009). "On the relationship between individual and population health". In: *Medicine, health care and philosophy* 12.3, 235–244.

Archer, E. and Archer, M. E. (2016). "Package 'rfPermute'". In: *R Project: Indianapolis, IN, USA*.

Arieno, A., Chan, A., and Destounis, S. V. (2019). "A review of the role of augmented intelligence in breast imaging: from automated breast density assessment to risk stratification". In: *American Journal of Roentgenology* 212.2, 259–270.

Arleo, A., Didimo, W., Liotta, G., and Montecchiani, F. (2017). "Large graph visualizations using a distributed computing platform". In: *Information Sciences* 381, 124–141.

Armstrong, C. W., McGregor, N. R., Lewis, D. P., Butt, H. L., and Gooley, P. R. (2015). "Metabolic profiling reveals anomalous energy metabolism and oxidative stress pathways in chronic fatigue syndrome patients". In: *Metabolomics* 11.6, 1626–1639.

Astin, J. A. (1998). "Why patients use alternative medicine: results of a national study". In: *jama* 279.19, 1548–1553.

Auffray, C., Chen, Z., and Hood, L. (2009). "Systems medicine: the future of medical genomics and healthcare". In: *Genome medicine* 1, 1–11.

Bach, B., Kerracher, N., Hall, K. W., Carpendale, S., Kennedy, J., and Henry Riche, N. (2016). "Telling stories about dynamic networks with graph comics". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3670–3682.

Bagassi, M. and Macchi, L. (2020). "Creative Problem Solving as Overcoming a Misunderstanding". In: *Frontiers in Education*. Vol. 5. Frontiers Media SA, 538202.

Bajzar, L., Morser, J., and Nesheim, M. (1996). "TAFI, or plasma procarboxypeptidase B, couples the coagulation and fibrinolytic cascades through the thrombin-thrombomodulin complex". In: *Journal of Biological Chemistry* 271.28, 16603–16608.

Baker, M. (2015). "Mathematical modelling of cytokine dynamics in arthritic disease". PhD thesis. University of Nottingham.

Barabasi, A.-L. and Oltvai, Z. N. (2004). "Network biology: understanding the cell's functional organization". In: *Nature reviews genetics* 5.2, 101–113.

Barber, S. (2018). "A truly 'transformative'MBA: Executive education for the fourth industrial revolution". In.

Barker, M., Chue Hong, N. P., Katz, D. S., Lamprecht, A.-L., Martinez-Ortiz, C., Psomopoulos, F., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., et al. (2022).

"Introducing the FAIR Principles for research software". In: *Scientific Data* 9.1, 622.
Bartoletti, I. (2019). "AI in healthcare: Ethical and privacy challenges". In: *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*. Springer, 7–10.
Barton, C., Chettipally, U., Zhou, Y., Jiang, Z., Lynn-Palevsky, A., Le, S., Calvert, J., and Das, R. (2019). "Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs". In: *Computers in biology and medicine* 109, 79–84.
Barupal, S. R., Soni, M. L., and Barupal, R. (2019). "Factors affecting mortality following necrotizing soft-tissue infections: randomized prospective study". In: *Journal of Emergencies, Trauma, and Shock* 12.2, 108.
Basadur, M., Finkbeiner, C. T., et al. (1983). "Identifying attitudinal factors related to ideation in creative problem solving". In.
Basadur, M. and Finkbeiner, C. T. (1985). "Measuring preference for ideation in creative problem-solving training". In: *The Journal of applied behavioral science* 21.1, 37–49.
Basadur, M. and Hausdorf, P. A. (1996). "Measuring divergent thinking attitudes related to creative problem solving and innovation management". In: *Creativity Research Journal* 9.1, 21–32.
Bayle, L., Chimalapati, S., Schoehn, G., Brown, J., Vernet, T., and Durmort, C. (2011). "Zinc uptake by Streptococcus pneumoniae depends on both AdcA and AdcAII and is essential for normal bacterial morphology and virulence". In: *Molecular microbiology* 82.4, 904–916.
Bechar, J., Sepehripour, S., Hardwicke, J., and Filobbos, G. (2017). "Laboratory risk indicator for necrotising fasciitis (LRINEC) score for the assessment of early necrotising fasciitis: a systematic review of the literature". In: *The Annals of The Royal College of Surgeons of England* 99.5, 341–346.
Bechlivanidis, C. and Lagnado, D. A. (2013). "Does the "why" tell us the "when"?" In: *Psychological Science* 24.8, 1563–1572.
Bellotti, D., Rowińska-Żyrek, M., and Remelli, M. (2021). "How zinc-binding systems, expressed by human pathogens, acquire zinc from the colonized host environment: a critical review on zincophores". In: *Current Medicinal Chemistry* 28.35, 7312.
Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, 289–300.
Bernard, G. R., Vincent, J.-L., Laterre, P.-F., LaRosa, S. P., Dhainaut, J.-F., Lopez-Rodriguez, A., Steingrub, J. S., Garber, G. E., Helterbrand, J. D., Ely, E. W., et al. (2001). "Efficacy and safety of recombinant human activated protein C for severe sepsis". In: *New England journal of medicine* 344.10, 699–709.
Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schäfer, H., Schütz, B., Spraul, M., and Tenori, L. (2009). "Individual human phenotypes in metabolic space and time". In: *Journal of proteome research* 8.9, 4264–4271.
Bernish, B. and Rijn, I. van de (1999). "Characterization of a two-component system in Streptococcus pyogenes which is involved in regulation of hyaluronic acid production". In: *Journal of Biological Chemistry* 274.8, 4786–4793.

Besag, J. (1975). "Statistical analysis of non-lattice data". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 24.3, 179–195.

Betts, H. C., Puttick, M. N., Clark, J. W., Williams, T. A., Donoghue, P. C., and Pisani, D. (2018). "Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin". In: *Nature ecology & evolution* 2.10, 1556–1562.

Birant, D. and Kut, A. (2007). "ST-DBSCAN: An algorithm for clustering spatial–temporal data". In: *Data & knowledge engineering* 60.1, 208–221.

Biron, D., Marché, L., Ponton, F., Loxdale, H., Galéotti, N., Renault, L., Joly, C., and Thomas, F. (2005). "Behavioural manipulation in a grasshopper harbouring hairworm: a proteomics approach". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1577, 2117–2126.

Bisno, A. L., Cockerill III, F. R., and Bermudez, C. T. (2000). "The initial outpatient-physician encounter in group A streptococcal necrotizing fasciitis". In: *Clinical infectious diseases* 31.2, 607–608.

Blanco, E., Schirmbeck, F., and Costa, C. (2019). "Vocational Education for the Industrial Revolution". In: *Smart Industry & Smart Education: Proceedings of the 15th International Conference on Remote Engineering and Virtual Instrumentation 15*. Springer, 649–658.

Bober, M., Enochsson, C., Collin, M., and Mörgelin, M. (2010). "Collagen VI is a subepithelial adhesive target for human respiratory tract pathogens". In: *Journal of innate immunity* 2.2, 160–166.

Bocking, N., Matsumoto, C.-l., Loewen, K., Teatero, S., Marchand-Austin, A., Gordon, J., Fittipaldi, N., and McGeer, A. (2017). "High incidence of invasive group A streptococcal infections in remote indigenous communities in Northwestern Ontario, Canada". In: *Open Forum Infectious Diseases*. Vol. 4. 1. Oxford University Press US, ofw243.

Bodansky, D. M., Begaj, I., Evison, F., Webber, M., Woodman, C. B., and Tucker, O. N. (2020). "A 16-year longitudinal cohort study of incidence and bacteriology of necrotising fasciitis in England". In: *World journal of surgery* 44, 2580–2591.

Bonilla, F. A. and Oettgen, H. C. (2010). "Adaptive immunity". In: *Journal of Allergy and Clinical Immunology* 125.2, S33–S40.

Bonne, S. L. and Kadri, S. S. (2017). "Evaluation and management of necrotizing soft tissue infections". In: *Infectious Disease Clinics* 31.3, 497–511.

Boorse, C. (1977). "Health as a theoretical concept". In: *Philosophy of science* 44.4, 542–573.

— (2014). "A second rebuttal on health". In: *Journal of medicine and philosophy* 39.6, 683–724.

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., and Pfister, H. (2013). "What makes a visualization memorable?" In: *IEEE transactions on visualization and computer graphics* 19.12, 2306–2315.

Borregaard, N., Sørensen, O. E., and Theilgaard-Mönch, K. (2007). "Neutrophil granules: a library of innate immunity proteins". In: *Trends in immunology* 28.8, 340–345.

Bos, N., Lefèvre, T., Jensen, A., and d'Ettorre, P. (2012). "Sick ants become unsociable". In: *Journal of evolutionary biology* 25.2, 342–351.

Bos, N., Sundström, L., Fuchs, S., and Freitak, D. (2015). "Ants medicate to fight disease". In: *Evolution* 69.11, 2979–2984.

Boyer, A., Vargas, F., Coste, F., Saubusse, E., Castaing, Y., Gbikpi-Benissan, G., Hilbert, G., and Gruson, D. (2009). "Influence of surgical treatment timing on mortality from necrotizing soft tissue infections requiring intensive care management". In: *Intensive care medicine* 35, 847–853.

Breiman, L. (2001). "Random forests". In: *Machine learning* 45.1, 5–32.

Brewer, W. F., Chinn, C. A., and Samarapungavan, A. (1998). "Explanation in scientists and children". In: *Minds and Machines* 8, 119–136.

Bruggeman, F. J. and Westerhoff, H. V. (2007). "The nature of systems biology". In: *TRENDS in Microbiology* 15.1, 45–50.

Brussow, H., Canchaya, C., and Hardt, W.-D. (2004). "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion". In: *Microbiology and molecular biology reviews* 68.3, 560–602.

Bruun, T., Kittang, B., De Hoog, B., Aardal, S., Flaatten, H., Langeland, N., Mylvaganam, H., Vindenes, H., and Skrede, S. (2013). "Necrotizing soft tissue infections caused by Streptococcus pyogenes and Streptococcus dysgalactiae subsp. equisimilis of groups C and G in western Norway". In: *Clinical Microbiology and Infection* 19.12, E545–E550.

Bruun, T., Oppegaard, O., Kittang, B. R., Mylvaganam, H., Langeland, N., and Skrede, S. (2016). "Etiology of cellulitis and clinical prediction of streptococcal disease: a prospective study". In: *Open forum infectious diseases*. Vol. 3. 1. Oxford University Press, ofv181.

Bruun, T., Rath, E., Madsen, M. B., Oppegaard, O., Nekludov, M., Arnell, P., Karlsson, Y., Babbar, A., Bergey, F., Itzek, A., et al. (2021). "Risk factors and predictors of mortality in streptococcal necrotizing soft-tissue infections: a multicenter prospective study". In: *Clinical Infectious Diseases* 72.2, 293–300.

Buehner, M. J. and Humphreys, G. R. (2009). "Causal binding of actions to their effects". In: *Psychological science* 20.10, 1221–1228.

— (2010). "Causal contraction: spatial binding in the perception of collision events". In: *Psychological Science* 21.1, 44–48.

Bulger, E., Maislin, G., Dankner, W., May, A., Edgar, R., and Shirvan, A. (2018). "682: Early plasma cytokine levels correlate with outcome in necrotizing soft tissue infections". In: *Critical Care Medicine* 46.1, 327.

Buschur, K. L., Chikina, M., and Benos, P. V. (2020). "Causal network perturbations for instance-specific analysis of single cell and disease samples". In: *Bioinformatics* 36.8, 2515–2521.

Bushel, P. R., Wolfinger, R. D., and Gibson, G. (2007). "Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes". In: *BMC Systems Biology* 1.1, 15.

Bynum, B. (2000). "Drapetomania". In: *Lancet (London, England)* 356.9241, 1615.

Cacciatore, S., Tenori, L., Luchinat, C., Bennett, P. R., and MacIntyre, D. A. (2017). "KODAMA: an R package for knowledge discovery and data mining". In: *Bioinformatics* 33.4, 621–623.

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., Steinhauser, D., Fernie, A. R., Willmitzer, L., and Hannah, M. A. (2011). "High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions". In: *The Plant Journal* 67.5, 869–884.

Camacho, D., De La Fuente, A., and Mendes, P. (2005). "The origin of correlations in metabolomics data". In: *Metabolomics* 1.1, 53–63.
Campbell, W. D., Lazoura, E., Okada, N., and Okada, H. (2002). "Inactivation of C3a and C5a octapeptides by carboxypeptidase R and carboxypeptidase N". In: *Microbiology and immunology* 46.2, 131–134.
Carlsson, I., Wendt, P. E., and Risberg, J. (2000). "On the neurobiology of creativity. Differences in frontal activity between high and low creative subjects". In: *Neuropsychologia* 38.6, 873–885.
Carr, A. J., Gibson, B., and Robinson, P. G. (2001). "Is quality of life determined by expectations or experience?" In: *Bmj* 322.7296, 1240–1243.
Carr, A. C., Shaw, G. M., Natarajan, R., et al. (2015). "Ascorbate-dependent vasopressor synthesis: a rationale for vitamin C administration in severe sepsis and septic shock?" In: *Critical Care* 19.1, 1–8.
Cartledge, D., Dove, L., Richardson, E., and Wilkie, R. (2020). "Necrotizing Soft Tissue Infections: Case Reports from the Patients Prospective". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 7–20.
Cassotti, M., Agogué, M., Camarda, A., Houdé, O., and Borst, G. (2016). "Inhibitory control as a core process of creative problem solving and idea generation from childhood to adulthood". In: *New directions for child and adolescent development* 2016.151, 61–72.
Castellino, F. J. (1995). "Human protein C and activated protein C: components of the human anticoagulation system". In: *Trends in cardiovascular medicine* 5.2, 55–62.
Chakravorty, S. S., Hales, D. N., and Herbert, J. I. (2008). "How problem-solving really works". In: *International Journal of Data Analysis Techniques and Strategies* 1.1, 44–59.
Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., and Tsaneva-Atanasova, K. (2019). "Artificial intelligence, bias and clinical safety". In: *BMJ Quality & Safety* 28.3, 231–237.
Chan, A. W., Mercier, P., Schiller, D., Bailey, R., Robbins, S., Eurich, D. T., Sawyer, M. B., and Broadhurst, D. (2016). "1 H-NMR urinary metabolomic profiling for diagnosis of gastric cancer". In: *British journal of cancer* 114.1, 59.
Chan, J. C., Chan, D. L., Diakos, C. I., Engel, A., Pavlakis, N., Gill, A., and Clarke, S. J. (2017). "The lymphocyte-to-monocyte ratio is a superior predictor of overall survival in comparison to established biomarkers of resectable colorectal cancer". In: *Annals of surgery* 265.3, 539.
Chan, T., Yaghoubian, A., Rosing, D., Kaji, A., and Virgilio, C. de (2008). "Low sensitivity of physical examination findings in necrotizing soft tissue infection is improved with laboratory values: a prospective study". In: *The American journal of surgery* 196.6, 926–930.
Chang, M. S., McNinch, J., Basu, R., and Simonet, S. (1994). "Cloning and characterization of the human neutrophil-activating peptide (ENA-78) gene." In: *Journal of Biological Chemistry* 269.41, 25277–25282.
Chaplin, D. D. (2010). "Overview of the immune response". In: *Journal of allergy and clinical immunology* 125.2, S3–S23.
Charon, R. (2008). *Narrative medicine: Honoring the stories of illness*. Oxford University Press.

Chatila, T. and Geha, R. S. (1993). "Signal transduction by microbial superantigens via MHC class II molecules". In: *Immunological reviews* 131.1, 43–59.
Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). "Analysis of microarray data using Z score transformation". In: *The Journal of molecular diagnostics* 5.2, 73–81.
Chella Krishnan, K., Mukundan, S., Alagarsamy, J., Hur, J., Nookala, S., Siemens, N., Svensson, M., Hyldegaard, O., Norrby-Teglund, A., and Kotb, M. (2016). "Genetic architecture of group A streptococcal necrotizing soft tissue infections in the mouse". In: *PLoS pathogens* 12.7, e1005732.
Chen, C.-Y., Luo, S.-C., Kuo, C.-F., Lin, Y.-S., Wu, J.-J., Lin, M. T., Liu, C.-C., Jeng, W.-Y., and Chuang, W.-J. (2003). "Maturation processing and characterization of streptopain". In: *Journal of Biological Chemistry* 278.19, 17336–17343.
Chen, T. and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4, 1–4.
Chen, X. and Jensen, P. E. (2008). "The role of B lymphocytes as antigen-presenting cells". In: *Archivum immunologiae et therapiae experimentalis* 56, 77–83.
Childers, B. J., Potyondy, L. D., Nachreiner, R., Rogers, F. R., Childers, E. R., Oberg, K. C., Hendricks, D. L., and Hardesty, R. A. (2002). "Necrotizing fasciitis: a fourteen-year retrospective study of 163 consecutive patients". In: *The American Surgeon* 68.2, 109–116.
Choi, B. C. and Anita, W. (2008). "Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 3. Discipline, inter-discipline distance, and selection of discipline". In: *Clinical and Investigative Medicine*, E41–E48.
Choi, B. C. and Pak, A. W. (2006). "Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness". In: *Clinical and investigative medicine* 29.6, 351.
— (2007). "Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. Promotors, barriers, and strategies of enhancement". In: *Clinical and Investigative Medicine*, E224–E232.
Choi, B. C., Pang, T., Lin, V., Puska, P., Sherman, G., Goddard, M., Ackland, M. J., Sainsbury, P., Stachenko, S., Morrison, H., et al. (2005). "Can scientists and policy makers work together?" In: *Journal of Epidemiology & Community Health* 59.8, 632–637.
Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S., and Xia, J. (2018). "MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis". In: *Nucleic acids research* 46.W1, W486–W494.
Clankie, S. (2013). "An overview of genericization in Linguistics". In: *Proceedings of the Second International Conference on Onomastics 'Name and Naming': Onomastics in Contemporary Public Space, Editura Mega, Editura Argonaut, Cluj-Napoca*.
Clark, C. J., Liu, B. S., Winegard, B. M., and Ditto, P. H. (2019). "Tribalism is human nature". In: *Current Directions in Psychological Science* 28.6, 587–592.

Clermont, G., Auffray, C., Moreau, Y., Rocke, D. M., Dalevi, D., Dubhashi, D., Marshall, D. R., Raasch, P., Dehne, F., Provero, P., et al. (2009). "Bridging the gap between systems biology and medicine". In: *Genome medicine* 1, 1–6.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289*.

Cocanour, C. S., Chang, P., Huston, J. M., Adams Jr, C. A., Diaz, J. J., Wessel, C. B., Falcione, B. A., Bauza, G. M., Forsythe, R. A., and Rosengart, M. R. (2017). "Management and novel adjuncts of necrotizing soft tissue infections". In: *Surgical infections* 18.3, 250–272.

Colaco, C. A., Bailey, C. R., Walker, K. B., and Keeble, J. (2013). "Heat shock proteins: stimulators of innate and acquired immunity". In: *BioMed research international* 2013.

Coleman, T. F. and Li, Y. (1996). "An interior trust region approach for nonlinear minimization subject to bounds". In: *SIAM Journal on optimization* 6.2, 418–445.

Collin, M. and Olsén, A. (2001). "EndoS, a novel secreted protein from Streptococcus pyogenes with endoglycosidase activity on human IgG". In: *The EMBO journal* 20.12, 3046–3055.

Comeau, A. M. and Krisch, H. M. (2005). "War is peace—dispatches from the bacterial and phage killing fields". In: *Current opinion in microbiology* 8.4, 488–494.

Commons, R. J., Smeesters, P. R., Proft, T., Fraser, J. D., Robins-Browne, R., and Curtis, N. (2014). "Streptococcal superantigens: categorization and clinical associations". In: *Trends in molecular medicine* 20.1, 48–62.

Consalvo, C. M. (1989). "Humor in management: No laughing matter". In.

Consortium, G. O. (2019). "The gene ontology resource: 20 years and still GOing strong". In: *Nucleic Acids Res* 47.D1, D330–D338.

Consortium, T. U. (2023). "UniProt: the Universal Protein knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1, D523–D531.

Consortium, U. (2019). "UniProt: a worldwide hub of protein knowledge". In: *Nucleic acids research* 47.D1, D506–D515.

Conway, E. M. (2012). "Thrombomodulin and its role in inflammation". In: *Seminars in immunopathology*. Vol. 34. Springer, 107–125.

Coughlin, S. R. (2005). "Protease-activated receptors in hemostasis, thrombosis and vascular biology". In: *Journal of Thrombosis and Haemostasis* 3.8, 1800–1814.

Coutinho, T. A. and Venter, S. N. (2009). "Pantoea ananatis: an unconventional plant pathogen". In: *Molecular plant pathology* 10.3, 325–335.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Cowie, J., Calveley, E., Bowers, G., and Bowers, J. (2018). "Evaluation of a digital consultation and self-care advice tool in primary care: a multi-methods study". In: *International journal of environmental research and public health* 15.5, 896.

Cropley, A. (2006). "In praise of convergent thinking". In: *Creativity research journal* 18.3, 391–404.

Csardi, G., Nepusz, T., et al. (2006). "The igraph software package for complex network research". In: *InterJournal, complex systems* 1695.5, 1–9.

Cummins, D. D., Lubart, T., Alksnis, O., and Rist, R. (1991). "Conditional reasoning and causation". In: *Memory & cognition* 19, 274–282.

Cummins, D. D. (1995). "Naive theories and causal deduction". In: *Memory & cognition* 23.5, 646–658.
Darenberg, J., Luca-Harari, B., Jasir, A., Sandgren, A., Pettersson, H., Schalén, C., Norgren, M., Romanus, V., Norrby-Teglund, A., and Normark, B. H. (2007). "Molecular and clinical characteristics of invasive group A streptococcal infection in Sweden". In: *Clinical infectious diseases* 45.4, 450–458.
Darwin, C. (2017). *On the Tendency of Species to Form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection by Charles Darwin-Delphi Classics (Illustrated)*. Vol. 8. Delphi Classics.
Darwin, C. and Wallace, A. (1858). "On The Tendency Of Species To Form Varieties; and On The Perpetuation Of Varieties & Species by Natural Means Of Selection." In: *Journal of the proceedings of the Linnean society*.
Das, D. K., Baker, M. G., and Venugopal, K. (2011). "Increasing incidence of necrotizing fasciitis in New Zealand: a nationwide study over the period 1990 to 2006". In: *Journal of Infection* 63.6, 429–433.
Davidson, R. and Harel, D. (1996). "Drawing graphs nicely using simulated annealing". In: *ACM Transactions on Graphics (TOG)* 15.4, 301–331.
Davies, H. D. (2001). "Flesh-eating disease: a note on necrotizing fasciitis". In: *Paediatrics & Child Health* 6.5, 243–247.
Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E. M., et al. (2020). "The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities". In: *Nucleic acids research* 48.D1, D606–D612.
De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. (2018). "Clinically applicable deep learning for diagnosis and referral in retinal disease". In: *Nature medicine* 24.9, 1342–1350.
De La Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). "Discovery of meaningful associations in genomic data using partial correlation coefficients". In: *Bioinformatics* 20.18, 3565–3574.
DeAngelis, P. L., Yang, N., and Weigel, P. H. (1994). "The Streptococcus pyogenes hyaluronan synthase: sequence comparison and conservation among various group A strains". In: *Biochemical and biophysical research communications* 199.1, 1–10.
DebRoy, S., Aliaga-Tobar, V., Galvez, G., Arora, S., Liang, X., Horstmann, N., Maracaja-Coutinho, V., Latorre, M., Hook, M., Flores, A. R., et al. (2021). "Genome-wide analysis of in vivo CcpA binding with and without its key co-factor HPr in the major human pathogen group A Streptococcus". In: *Molecular Microbiology* 115.6, 1207–1228.
Denzer, L., Schroten, H., and Schwerk, C. (2020). "From gene to protein—How bacterial virulence factors manipulate host gene expression during infection". In: *International Journal of Molecular Sciences* 21.10, 3730.
Di Sanzo, M., Quaresima, B., Biamonte, F., Palmieri, C., and Faniello, M. C. (2020). "FTH1 pseudogenes in cancer and cell metabolism". In: *Cells* 9.12, 2554.
Dinarello, C. A. (2000). "Proinflammatory cytokines". In: *Chest* 118.2, 503–508.
Dinkla, K., Rohde, M., Jansen, W. T., Kaplan, E. L., Chhatwal, G. S., Talay, S. R., et al. (2003). "Rheumatic fever–associated Streptococcus pyogenes isolates aggregate collagen". In: *The Journal of clinical investigation* 111.12, 1905–1912.

Dinkla, K., Sastalla, I., Godehardt, A. W., Janze, N., Chhatwal, G. S., Rohde, M., and Medina, E. (2007). "Upregulation of capsule enables Streptococcus pyogenes to evade immune recognition by antigen-specific antibodies directed to the G-related α2-macroglobulin-binding protein GRAB located on the bacterial surface". In: *Microbes and infection* 9.8, 922–931.

Domingos, P. (1999). "The role of Occam's razor in knowledge discovery". In: *Data mining and knowledge discovery* 3, 409–425.

Doquire, G., Verleysen, M., et al. (2012). "A Comparison of Multivariate Mutual Information Estimators for Feature Selection." In: *ICPRAM (1)*, 176–185.

Doron, S. and Gorbach, S. L. (2008). "Bacterial infections: overview". In: *International Encyclopedia of Public Health*, 273.

Draghi, M. (2020). "We face a war against coronavirus and must mobilise accordingly". In: *Financial Times* 25.3.

Du Preez, P. (2012). "The human right to education, the ethical responsibility of curriculum, and the irony in 'safe spaces'". In: *Safe Spaces*. Brill, 51–62.

Dubois, C., Marcé, D., Faivre, V., Lukaszewicz, A.-C., Junot, C., Fenaille, F., Simon, S., Becher, F., Morel, N., and Payen, D. (2019). "High plasma level of S100A8/S100A9 and S100A12 at admission indicates a higher risk of death in septic shock patients". In: *Scientific reports* 9.1, 1–7.

Dubois, C., Payen, D., Simon, S., Junot, C., Fenaille, F., Morel, N., and Becher, F. (2020). "Top-down and bottom-up proteomics of circulating S100A8/S100A9 in plasma of septic shock patients". In: *Journal of proteome research* 19.2, 914–925.

Dudley, J. T., Listgarten, J., Stegle, O., Brenner, S. E., and Parts, L. (2014). "Personalized medicine: from genotypes, molecular phenotypes and the quantified self, towards improved medicine". In: *Pacific symposium on biocomputing co-chairs*. World Scientific, 342–346.

Dyer, M. D., Neff, C., Dufford, M., Rivera, C. G., Shattuck, D., Bassaganya-Riera, J., Murali, T., and Sobral, B. W. (2010). "The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis". In: *PloS one* 5.8, e12089.

Dyson, A. (2016). "Interactive visualization for interdisciplinary research". In: Visualization and Data Analysis.

Eberle, R. F. (1972). "Developing imagination through scamper." In: *Journal of Creative Behavior*.

Edoardo, S., HWB, H. M., et al. (2020). "Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models". In: *Scientific Reports (Nature Publisher Group)* 10.1.

Egesten, A., Eliasson, M., Johansson, H. M., Olin, A. I., Mörgelin, M., Mueller, A., Pease, J. E., Frick, I.-M., and Björck, L. (2007). "The CXC chemokine MIG/CXCL9 is important in innate immunity against Streptococcus pyogenes". In: *The Journal of infectious diseases* 195.5, 684–693.

Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S., Greiner, R., Wishart, D. S., and Baracos, V. E. (2011). "Learning to predict cancer-associated skeletal muscle wasting from 1 H-NMR profiles of urinary metabolites". In: *Metabolomics* 7.1, 25–34.

Elefsinioti, A., Bellaire, T., Wang, A., Quast, K., Seidel, H., Braxenthaler, M., Goeller, G., Christianson, A., Henderson, D., and Reischl, J. (2016). "Key factors for suc-

cessful data integration in biomarker research". In: *Nature Reviews Drug Discovery* 15.6, 369–370.

Elliott, D., Kufera, J. A., and Myers, R. A. (2000). "The microbiology of necrotizing soft tissue infections". In: *The American journal of surgery* 179.5, 361–366.

Elphick, G. F., Sarangi, P. P., Hyun, Y.-M., Hollenbaugh, J. A., Ayala, A., Biffl, W. L., Chung, H.-L., Rezaie, A. R., McGrath, J. L., Topham, D. J., et al. (2009). "Recombinant human activated protein C inhibits integrin-mediated neutrophil migration". In: *Blood, The Journal of the American Society of Hematology* 113.17, 4078–4085.

Emgaard, J., Bergsten, H., McCormick, J. K., Barrantes, I., Skrede, S., Sandberg, J. K., and Norrby-Teglund, A. (2019). "MAIT cells are major contributors to the cytokine response in group A streptococcal toxic shock syndrome". In: *Proceedings of the National Academy of Sciences* 116.51, 25923–25931.

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., et al. (2019). "NMR spectroscopy for metabolomics research". In: *Metabolites* 9.7, 123.

Engel, H., Gutierrez-Fernandez, J., Fluckiger, C., Martinez-Ripoll, M., Muhlemann, K., Hermoso, J. A., Hilty, M., and Hathaway, L. J. (2013). "Heteroresistance to fosfomycin is predominant in Streptococcus pneumoniae and depends on the murA1 gene". In: *Antimicrobial agents and chemotherapy* 57.6, 2801–2808.

Eraso, J. M., Olsen, R. J., Beres, S. B., Kachroo, P., Porter, A. R., Nasser, W., Bernard, P. E., DeLeo, F. R., and Musser, J. M. (2016). "Genomic Landscape of Intrahost Variation in Group A Streptococcus: Repeated and Abundant Mutational Inactivation of the fabT Gene Encoding a Regulator of Fatty Acid Synthesis". In: *Infection and Immunity* 84.12. Ed. by V. B. Young, 3268–3281. DOI: 10.1128/IAI.00608-16. URL: https://journals.asm.org/doi/10.1128/IAI.00608-16.

Erdel, M., Laich, A., Utermann, G., Werner, E., and Werner-Felmayer, G. (1998). "The human gene encoding SCYB9B, a putative novel CXC chemokine, maps to human chromosome 4q21 like the closely related genes for MIG (SCYB9) and INP10 (SCYB10)". In: *Cytogenetic and Genome Research* 81.3/4, 271.

Erichsen Andersson, A., Egerod, I., Knudsen, V. E., and Fagerdahl, A.-M. (2018). "Signs, symptoms and diagnosis of necrotizing fasciitis experienced by survivors and family: a qualitative Nordic multi-center study". In: *BMC infectious diseases* 18, 1–9.

Eron, L. J., Lipsky, B. A., Low, D. E., Nathwani, D., Tice, A. D., and Volturo, G. A. (2003). "Managing skin and soft tissue infections: expert panel recommendations on key decision points". In: *Journal of Antimicrobial Chemotherapy* 52.suppl_1, i3–i17.

Esmon, C. T., Esmon, N., and Harris, K. (1982). "Complex formation between thrombin and thrombomodulin inhibits both thrombin-catalyzed fibrin formation and factor V activation." In: *Journal of Biological Chemistry* 257.14, 7944–7947.

Esmon, C. (2005). "Do-all receptor takes on coagulation, inflammation". In: *Nature medicine* 11.5, 475–477.

Esmon, C. T. (1993). "Molecular events that control the protein C anticoagulant pathway". In: *Thrombosis and haemostasis* 70.07, 029–035.

— (2005). "The interactions between inflammation and coagulation". In: *British journal of haematology* 131.4, 417–430.

Esmon, N., Carroll, R., and Esmon, C. (1983). "Thrombomodulin blocks the ability of thrombin to activate platelets." In: *Journal of Biological Chemistry* 258.20, 12238–12242.

Fahrmann, J. F., Kim, K., DeFelice, B. C., Taylor, S. L., Gandara, D. R., Yoneda, K. Y., Cooke, D. T., Fiehn, O., Kelly, K., and Miyamoto, S. (2015). "Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer". In: *Cancer Epidemiology and Prevention Biomarkers* 24.11, 1716–1723.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles". In: *PLoS biology* 5.1, e8.

Fang, Y., Li, C., Shao, R., Yu, H., and Zhang, Q. (2018). "The role of biomarkers of endothelial activation in predicting morbidity and mortality in patients with severe sepsis and septic shock in intensive care: a prospective observational study". In: *Thrombosis Research* 171, 149–154.

Fauchald, S. K. and Smith, D. (2005). "Transdisciplinary research partnerships: Making research happen!" In: *Nursing Economics* 23.3, 131.

Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, 861–874.

Fedor-Freybergh, P. G. (1999). "Psychoimmuno-neuroendocrinology: An integrative approach to modern philosophy in medicine and psychology". In: *Neuroendocrinology Letters* 20, 205–220.

Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M., and Reinhardt, J. D. (2013). "Stable graphical model estimation with random forests for discrete, continuous, and mixed variables". In: *Computational Statistics & Data Analysis* 64, 132–152.

Fernando, S. M., Tran, A., Cheng, W., Rochwerg, B., Kyeremanteng, K., Seely, A. J., Inaba, K., and Perry, J. J. (2019). "Necrotizing soft tissue infection: diagnostic accuracy of physical examination, imaging, and LRINEC score: a systematic review and meta-analysis". In: *Annals of surgery* 269.1, 58–65.

Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., Neuper, C., Ebner, F., and Neubauer, A. C. (2009). "The creative brain: Investigation of brain activity during creative problem solving by means of EEG and fMRI". In: *Human brain mapping* 30.3, 734–748.

Fitzgerald, J., Rich, C., Zhou, F. H., and Hansen, U. (2008). "Three novel collagen VI chains, α4 (VI), α5 (VI), and α6 (VI)". In: *Journal of Biological Chemistry* 283.29, 20170–20180.

Flanigan, L. K. (2021). "I do not have stigma towards people with ADHD (but I do think they're lazy): Using education and experience to reduce negative attitudes towards ADHD". In.

Fleenor, J. W. (2006). "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies and nations". In: *Personnel Psychology* 59.4, 982.

Fleming, N. (2018). "Computer-calculated compounds". In: *Nature* 557.7707, S55–7.

Flensner, K. K. and Von der Lippe, M. (2019). "Being safe from what and safe for whom? A critical discussion of the conceptual metaphor of 'safe space'". In: *Intercultural Education* 30.3, 275–288.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.

Forina, M., Armanino, C., Castino, M., and Ubigli, M. (1986). "Multivariate data analysis as a discriminating method of the origin of wines". In: *Vitis* 25.3, 189–201.

Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983). "Classification of olive oils from their fatty acid composition". In: *Food research and data analysis: proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway/edited by H. Martens and H. Russwurm, Jr*. London: Applied Science Publishers, 1983.

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., et al. (2018). "Species-level functional profiling of metagenomes and metatranscriptomes". In: *Nature methods* 15.11, 962–968.

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., et al. (2019). "Gut microbiome structure and metabolic activity in inflammatory bowel disease". In: *Nature microbiology* 4.2, 293.

Freeman, R. B. and Huang, W. (2014). "Collaboration: Strength in diversity". In: *Nature* 513.7518, 305–305.

Freund, Y. and Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1, 119–139.

Frick, I.-M., Shannon, O., Neumann, A., Karlsson, C., Wikström, M., and Björck, L. (2018). "Streptococcal inhibitor of complement (SIC) modulates fibrinolysis and enhances bacterial survival within fibrin clots". In: *Journal of Biological Chemistry* 293.35, 13578–13591.

Friedman, J. and Alm, E. J. (2012). "Inferring correlation networks from genomic survey data". In: *PLoS computational biology* 8.9, e1002687.

FRooN, A. H., Bemelmans, M. H., GREvE, J. W., VAN DER LINDEN, C. J., and Buurman, W. A. (1994). "Increased plasma concentrations of soluble tumor necrosis factor receptors in sepsis syndrome: correlation with plasma creatinine values". In: *Critical care medicine* 22.5, 803–809.

Fuhrman, J. A. (1999). "Marine viruses and their biogeochemical and ecological effects". In: *Nature* 399.6736, 541–548.

Fukushima, K. (1975). "Cognitron: A self-organizing multilayered neural network". In: *Biological cybernetics* 20.3-4, 121–136.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). "A census of human cancer genes". In: *Nature reviews cancer* 4.3, 177–183.

Gaardlund, B. (2006). "Activated protein C (Xigris) treatment in sepsis: A drug in trouble". In: *Acta anaesthesiologica scandinavica* 50.8, 907–910.

Galperin, G. (2003). "Playing pool with π (the number π from a billiard point of view)". In: *Regular and chaotic dynamics* 8.4, 375–394.

Ganna, A., Salihovic, S., Sundström, J., Broeckling, C. D., Hedman, A. K., Magnusson, P. K., Pedersen, N. L., Larsson, A., Siegbahn, A., Zilmer, M., et al. (2014). "Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease". In: *PLoS genetics* 10.12, e1004801.

Garriga, H., Von Krogh, G., and Spaeth, S. (2013). "How constraints and knowledge impact open innovation". In: *Strategic Management Journal* 34.9, 1134–1144.

Gautam, N., Maria Olofsson, A., Herwald, H., Iversen, L. F., Lundgren-AAkerlund, E., Hedqvist, P., Arfors, K.-E., Flodgaard, H., and Lindbom, L. (2001). "Heparin-binding protein (HBP/CAP37): a missing link in neutrophil-evoked alteration of vascular permeability". In: *Nature medicine* 7.10, 1123–1127.

Gelfand, I. M. and Yaglom, A. M. (1957). "Calculation of amount of information about a random function contained in another such function". In: *American Mathematical Society Translations* 2.12, 199–246.

Gelfert, A. (2018). "Models in search of targets: exploratory modelling and the case of turing patterns". In: *Philosophy of Science: Between the Natural Sciences, the Social Sciences, and the Humanities*, 245–269.

Gesiarz, F., Cahill, D., and Sharot, T. (2019). "Evidence accumulation is biased by motivation: A computational account". In: *PLoS computational biology* 15.6, e1007089.

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y., and Kitano, H. (2011). "Software for systems biology: from tools to integrated platforms". In: *Nature Reviews Genetics* 12.12, 821–832.

Ghosh, S. and Henderson, S. G. (2003). "Behavior of the NORTA method for correlated random vector generation as the dimension increases". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 13.3, 276–294.

Gibson, H., Faith, J., and Vickers, P. (2013). "A survey of two-dimensional graph layout techniques for information visualisation". In: *Information visualization* 12.3-4, 324–357.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022). "The reactome pathway knowledgebase 2022". In: *Nucleic acids research* 50.D1, D687–D692.

Gillespie, T., Boczkowski, P. J., and Foot, K. A. (2014). *Media technologies: Essays on communication, materiality, and society*. MIT Press.

Gilliet, M., Cao, W., and Liu, Y.-J. (2008). "Plasmacytoid dendritic cells: sensing nucleic acids in viral infection and autoimmune diseases". In: *Nature Reviews Immunology* 8.8, 594–606.

Giuliano, A., Lewis Jr, F., Hadley, K., and Blaisdell, F. W. (1977). "Bacteriology of necrotizing fasciitis". In: *The American Journal of Surgery* 134.1, 52–57.

Glass, G., Sheil, F., Ruston, J., and Butler, P. (2015). "Necrotising soft tissue infection in a UK metropolitan population". In: *The Annals of The Royal College of Surgeons of England* 97.1, 46–51.

Glavey, S. V., Naba, A., Manier, S., Clauser, K., Tahri, S., Park, J., Reagan, M. R., Moschetta, M., Mishima, Y., Gambella, M., et al. (2017). "Proteomic characterization of human multiple myeloma bone marrow extracellular matrix". In: *Leukemia* 31.11, 2426–2434.

Goh, T., Goh, L., Ang, C., and Wong, C. (2014). "Early diagnosis of necrotizing fasciitis". In: *Journal of British Surgery* 101.1, e119–e125.

Goldmann, O., Köckritz-Blickwede, M. von, Höltje, C., Chhatwal, G. S., Geffers, R., and Medina, E. (2007). "Transcriptome analysis of murine macrophages in response to infection with Streptococcus pyogenes reveals an unusual activation program". In: *Infection and immunity* 75.8, 4148–4157.

Goldstein, E. J., Anaya, D. A., and Dellinger, E. P. (2007). "Necrotizing soft-tissue infection: diagnosis and management". In: *Clinical Infectious Diseases* 44.5, 705–710.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). "Data integration in the era of omics: current and future challenges". In: *BMC systems biology* 8.2, 1–10.

Goosens, W. K. (1980). "Values, health, and medicine". In: *Philosophy of Science* 47.1, 100–115.

Göttlich, S., Herty, M., and Klar, A. (2005). "Network models for supply chains". In: *Communications in Mathematical Sciences* 3.4, 545–559.

Gottlieb, M., Long, B., and Koyfman, A. (2018). "The evaluation and management of toxic shock syndrome in the emergency department: a review of the literature". In: *The Journal of emergency medicine* 54.6, 807–814.

Graff, H. J. (2016). "The "problem" of interdisciplinarity in theory, practice, and history". In: *Social Science History* 40.4, 775–803.

Gratz, N., Siller, M., Schaljo, B., Pirzada, Z. A., Gattermeier, I., Vojtek, I., Kirschning, C. J., Wagner, H., Akira, S., Charpentier, E., et al. (2008). "Group A streptococcus activates type I interferon production and MyD88-dependent signaling without involvement of TLR2, TLR4, and TLR9". In: *Journal of Biological Chemistry* 283.29, 19879–19887.

Green, M. C. and Brock, T. C. (2000). "The role of transportation in the persuasiveness of public narratives." In: *Journal of personality and social psychology* 79.5, 701.

Greenspan, D. S., Byers, M. G., Eddy, R. L., Cheng, W., Jani-Sait, S., and Shows, T. B. (1992). "Human collagen gene COL5A1 maps to the q34. 2→ q34. 3 region of chromosome 9, near the locus for nail-patella syndrome". In: *Genomics* 12.4, 836–837.

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., and Hermjakob, H. (2020). "ReactomeGSA-efficient multi-omics comparative pathway analysis". In: *Molecular & Cellular Proteomics* 19.12, 2115–2125.

Gross, D. J. (1996). "The role of symmetry in fundamental physics". In: *Proceedings of the National Academy of Sciences* 93.25, 14256–14259.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). "Recent advances in convolutional neural networks". In: *Pattern recognition* 77, 354–377.

Gu, X., Ritter, S. M., Delfmann, L. R., and Dijksterhuis, A. (2022). "Stimulating creativity: Examining the effectiveness of four cognitive-based creativity training techniques". In: *The Journal of Creative Behavior* 56.3, 312–327.

Guclu, E., Durmaz, Y., and Karabay, O. (2013). "Effect of severe sepsis on platelet count and their indices". In: *African health sciences* 13.2, 333–338.

Gunderson, C. G., Cherry, B. M., and Fisher, A. (2018). "Do patients with cellulitis need to be hospitalized? A systematic review and meta-analysis of mortality rates of inpatients with cellulitis". In: *Journal of General Internal Medicine* 33, 1553–1560.

Guo, X., Liu, Y., Li, D., and Li, Y. (2019). "Plasma Thrombomodulin Levels are Associated with Endothelial Injury in Patients with Bacterial Infections." In: *Clinical Laboratory* 65.9.

Guo, Y., Hastie, T., and Tibshirani, R. (2006). "Regularized linear discriminant analysis and its application in microarrays". In: *Biostatistics* 8.1, 86–100.

Gustafsson, M., Nestor, C. E., Zhang, H., Barabási, A.-L., Baranzini, S., Brunak, S., Chung, K. F., Federoff, H. J., Gavin, A.-C., Meehan, R. R., et al. (2014). "Modules, networks and systems medicine for understanding disease and aiding diagnosis". In: *Genome medicine* 6, 1–11.

Hadeed, G. J., Smith, J., O'Keeffe, T., Kulvatunyou, N., Wynne, J. L., Joseph, B., Friese, R. S., Wachtel, T. L., Rhee, P. M., El-Menyar, A., et al. (2016). "Early surgical intervention and its impact on patients presenting with necrotizing soft tissue infections: a single academic center experience". In: *Journal of emergencies, trauma, and shock* 9.1, 22.

Haelewaters, D., Hofmann, T. A., and Romero-Olivares, A. L. (2021). "Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South". In: *PLoS Computational Biology* 17.8, e1009277.

Haglund, C. M. and Welch, M. D. (2011). "Pathogens and polymers: microbe–host interactions illuminate the cytoskeleton". In: *Journal of Cell Biology* 195.1, 7–17.

Hamburg, M. A. and Collins, F. S. (2010). "The path to personalized medicine". In: *New England Journal of Medicine* 363.4, 301–304.

Han, C., Zhong, J., Hu, J., Liu, H., Liu, R., and Ling, F. (2020). "Single-Sample Node Entropy for Molecular Transition in Pre-deterioration Stage of Cancer". In: *Frontiers in Bioengineering and Biotechnology* 8, 809.

Hansen, M. B., Rasmussen, L. S., Garred, P., Pilely, K., Wahl, A. M., Perner, A., Madsen, M. B., Hedegaard, E. R., Simonsen, U., and Hyldegaard, O. (2018). "Associations of plasma nitrite, L-arginine and asymmetric dimethylarginine with morbidity and mortality in patients with necrotizing soft tissue infections". In: *Shock: Injury, Inflammation, and Sepsis: Laboratory and Clinical Approaches* 49.6, 667–674.

Hansen, M. B., Rasmussen, L. S., Garred, P., Bidstrup, D., Madsen, M. B., and Hyldegaard, O. (2016). "Pentraxin-3 as a marker of disease severity and risk of death in patients with necrotizing soft tissue infections: a nationwide, prospective, observational study". In: *Critical Care* 20, 1–11.

Hansen, M. B., Rasmussen, L. S., Svensson, M., Chakrakodi, B., Bruun, T., Madsen, M. B., Perner, A., Garred, P., Hyldegaard, O., Norrby-Teglund, A., et al. (2017). "Association between cytokine response, the LRINEC score and outcome in patients with necrotising soft tissue infection: a multicentre, prospective study". In: *Scientific Reports* 7.1, 42179.

Hansen, M. B., Simonsen, U., Garred, P., and Hyldegaard, O. (2015). "Biomarkers of necrotising soft tissue infections: aspects of the innate immune response and effects of hyperbaric oxygenation—the protocol of the prospective cohort BIONEC study". In: *BMJ open* 5.5, e006995.

Hardin, J., Garcia, S. R., and Golan, D. (2013). "A method for generating realistic correlation matrices". In: *The Annals of Applied Statistics*, 1733–1762.

Hartigan, J. A. and Wong, M. A. (1979). "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the royal statistical society. series c (applied statistics)* 28.1, 100–108.

Hasin, Y., Seldin, M., and Lusis, A. (2017). "Multi-omics approaches to disease". In: *Genome biology* 18.1, 1–15.

Hathroubi, S., Zerebinski, J., and Ottemann, K. M. (2018). "Helicobacter pylori biofilm involves a multigene stress-biased response, including a structural role for flagella". In: *MBio* 9.5, 10–1128.

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., and O'Donovan, C. (2019). "MetaboLights: a resource evolving in response to the needs of its scientific community". In: *Nucleic acids research*.

Hay, S. I., George, D. B., Moyes, C. L., and Brownstein, J. S. (2013). "Big data opportunities for global infectious disease surveillance". In: *PLoS medicine* 10.4, e1001413.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Hedetoft, M., Garred, P., Madsen, M. B., and Hyldegaard, O. (2021). "Hyperbaric oxygen treatment is associated with a decrease in cytokine levels in patients with necrotizing soft-tissue infection". In: *Physiological Reports* 9.6, e14757.

Hedetoft, M., Madsen, M. B., Madsen, L. B., and Hyldegaard, O. (2020). "Incidence, comorbidity and mortality in patients with necrotising soft-tissue infections, 2005–2018: a Danish nationwide register-based cohort study". In: *BMJ open* 10.10, e041302.

Hendrycks, D. and Gimpel, K. (2016). "Gaussian error linear units (gelus)". In: *arXiv preprint arXiv:1606.08415*.

Hertzén, E., Johansson, L., Wallin, R., Schmidt, H., Kroll, M., Rehn, A. P., Kotb, M., Mörgelin, M., and Norrby-Teglund, A. (2010). "M1 protein-dependent intracellular trafficking promotes persistence and replication of Streptococcus pyogenes in macrophages". In: *Journal of innate immunity* 2.6, 534–545.

Herwald, H., Cramer, H., Mörgelin, M., Russell, W., Sollenberg, U., Norrby-Teglund, A., Flodgaard, H., Lindbom, L., and Björck, L. (2004). "M protein, a classical bacterial virulence determinant, forms complexes with fibrinogen that induce vascular leakage". In: *Cell* 116.3, 367–379.

Hesslow, G. (1993). "Do we need a concept of disease?" In: *Theoretical medicine* 14, 1–14.

Hilvo, M., Gade, S., Hyötyläinen, T., Nekljudova, V., Seppänen-Laakso, T., Sysi-Aho, M., Untch, M., Huober, J., Minckwitz, G. von, Denkert, C., et al. (2014). "Monounsaturated fatty acids in serum triacylglycerols are associated with response to neoadjuvant chemotherapy in breast cancer patients". In: *International journal of cancer* 134.7, 1725–1733.

Holub, M., Lawrence, D. A., Andersen, N., Davidová, A., Beran, O., Marešová, V., and Chalupa, P. (2013). "Cytokines and chemokines as biomarkers of community-acquired bacterial infection". In: *Mediators of inflammation* 2013.

Holzinger, A., Haibe-Kains, B., and Jurisica, I. (2019). "Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data". In: *European Journal of Nuclear Medicine and Molecular Imaging* 46, 2722–2730.

Hood, L. and Friend, S. H. (2011). "Predictive, personalized, preventive, participatory (P4) cancer medicine". In: *Nature reviews Clinical oncology* 8.3, 184–187.

Hsiao, C.-T., Chang, C.-P., Huang, T.-Y., Chen, Y.-C., and Fann, W.-C. (2020). "Prospective validation of the laboratory risk indicator for necrotizing fasciitis (LRINEC) score for necrotizing fasciitis of the extremities". In: *PLoS One* 15.1, e0227748.

Hua, C., Urbina, T., Bosc, R., Parks, T., Sriskandan, S., Prost, N. de, and Chosidow, O. (2022). "Necrotising soft-tissue infections". In: *The Lancet Infectious Diseases*.

Huang, K.-F., Hung, M.-H., Lin, Y.-S., Lu, C.-L., Liu, C., Chen, C.-C., and Lee, Y.-H. (2011). "Independent predictors of mortality for necrotizing fasciitis: a retrospective analysis in a single institution". In: *Journal of Trauma and Acute Care Surgery* 71.2, 467–473.

Huston, D. P. (1997). "The biology of the immune system". In: *Jama* 278.22, 1804–1814.

Iba, T., Hagiwara, A., Saitoh, D., Anan, H., Ueki, Y., Sato, K., and Gando, S. (2017). "Effects of combination therapy using antithrombin and thrombomodulin for sepsis-associated disseminated intravascular coagulation". In: *Annals of Intensive Care* 7.1, 1–10.

Iba, T., Yagi, Y., Kidokoro, A., Fukunaga, M., and Fukunaga, T. (1995). "Increased plasma levels of soluble thrombomodulin in patients with sepsis and organ failure". In: *Surgery today* 25, 585–590.

Ignatius, A., Schoengraf, P., Kreja, L., Liedert, A., Recknagel, S., Kandert, S., Brenner, R. E., Schneider, M., Lambris, J. D., and Huber-Lang, M. (2011). "Complement C3a and C5a modulate osteoclast formation and inflammatory response of osteoblasts in synergism with IL-1β". In: *Journal of cellular biochemistry* 112.9, 2594–2605.

Isermann, B., Hendrickson, S. B., Zogg, M., Wing, M., Cummiskey, M., Kisanuki, Y. Y., Yanagisawa, M., Weiler, H., et al. (2001). "Endothelium-specific loss of murine thrombomodulin disrupts the protein C anticoagulant pathway and causes juvenile-onset thrombosis". In: *The Journal of clinical investigation* 108.4, 537–546.

Ito, T., Chiba, T., and Yoshida, M. (2001). "Exploring the protein interactome using comprehensive two-hybrid projects". In: *Trends in biotechnology* 19, 23–27.

Jacobs, A., Kilb, D., and Kent, G. (2008). "3-D interdisciplinary visualization: Tools for scientific analysis and communication". In: *Seismological Research Letters* 79.6, 867–876.

Jahagirdar, S., Morris, L., Benis, N., Oppegaard, O., Svenson, M., Hyldegaard, O., Skrede, S., Norrby-Teglund, A., Martins dos Santos, V. A., and Saccenti, E. (2022). "Analysis of host-pathogen gene association networks reveals patient-specific response to streptococcal and polymicrobial necrotising soft tissue infections". In: *BMC medicine* 20.1, 1–18.

Jahagirdar, S. and Saccenti, E. (2020a). "Evaluation of single sample network inference methods for metabolomics-based systems medicine". In: *Journal of Proteome Research* 20.1, 932–949.

— (2020b). "On the Use of Correlation and MI as a Measure of Metabolite—Metabolite Association for Network Differential Connectivity Analysis". In: *Metabolites* 10.4, 171.

Jahagirdar, S., Suarez-Diez, M., and Saccenti, E. (2019). "Simulation and Reconstruction of Metabolite–Metabolite Association Networks Using a Metabolic Dynamic Model and Correlation Based Algorithms". In: *Journal of proteome research* 18.3, 1099–1113.

Jin, H., Agarwal, S., Agarwal, S., and Pancholi, V. (2011). "Surface export of GAPDH/SDH, a glycolytic enzyme, is essential for Streptococcus pyogenes virulence". In: *MBio* 2.3, e00068–11.

Jinawath, N., Bunbanjerdsuk, S., Chayanupatkul, M., Ngamphaiboon, N., Asavapanumas, N., Svasti, J., and Charoensawan, V. (2016). "Bridging the gap between clinicians and systems biologists: from network biology to translational biomedical research". In: *Journal of translational medicine* 14.1, 1–13.

Johansson, L., Linnér, A., Sundén-Cullberg, J., Haggar, A., Herwald, H., Loré, K., Treutiger, C.-J., and Norrby-Teglund, A. (2009). "Neutrophil-derived hyperresistinemia in severe acute streptococcal infections". In: *The Journal of Immunology* 183.6, 4047–4054.

Johansson, L. and Norrby-Teglund, A. (2012). "Immunopathogenesis of streptococcal deep tissue infections". In: *Host-pathogen interactions in streptococcal diseases*, 173–188.

Johansson, L., Snäll, J., Sendi, P., Linnér, A., Thulin, P., Linder, A., Treutiger, C.-J., and Norrby-Teglund, A. (2014). "HMGB1 in severe soft tissue infections caused by Streptococcus pyogenes". In: *Frontiers in cellular and infection microbiology* 4, 4.

Johansson, L., Thulin, P., Low, D. E., and Norrby-Teglund, A. (2010). "Getting under the skin: the immunopathogenesis of Streptococcus pyogenes deep tissue infections". In: *Clinical Infectious Diseases* 51.1, 58–65.

Joyce, D. E. and Grinnell, B. W. (2002). "Recombinant human activated protein C attenuates the inflammatory response in endothelium and monocytes by modulating nuclear factor-κB". In: *Critical care medicine* 30.5, S288–S293.

Juengst, E., McGowan, M. L., Fishman, J. R., and Settersten Jr, R. A. (2016). "From "personalized" to "precision" medicine: the ethical and social implications of rhetorical reform in genomic medicine". In: *Hastings Center Report* 46.5, 21–33.

Kachroo, P., Eraso, J. M., Olsen, R. J., Zhu, L., Kubiak, S. L., Pruitt, L., Yerramilli, P., Cantu, C. C., Ojeda Saavedra, M., Pensar, J., et al. (2020). "New pathogenesis mechanisms and translational leads identified by multidimensional analysis of necrotizing myositis in primates". In: *MBio* 11.1, e03363–19.

Kahn, F., Mörgelin, M., Shannon, O., Norrby-Teglund, A., Herwald, H., Olin, A. I., and Björck, L. (2008). "Antibodies against a surface protein of Streptococcus pyogenes promote a pathological inflammatory response". In: *PLoS pathogens* 4.9, e1000149.

Kalmady, S. V., Greiner, R., Agrawal, R., Shivakumar, V., Narayanaswamy, J. C., Brown, M. R., Greenshaw, A. J., Dursun, S. M., and Venkatasubramanian, G. (2019). "Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning". In: *npj Schizophrenia* 5.1, 2.

Kapiotis, S., Besemer, J., Bevec, D., Valent, P., Bettelheim, P., Lechner, K., and Speiser, W. (1991). "Interleukin-4 counteracts pyrogen-induced downregulation of thrombomodulin in cultured human vascular endothelial cells". In.

Katz, S., Suijker, J., Hardt, C., Madsen, M. B., Meij-de Vries, A., Pijpe, A., Skrede, S., Hyldegaard, O., Solligård, E., Norrby-Teglund, A., et al. (2022). "Decision support system and outcome prediction in a cohort of patients with necrotizing soft-tissue infections". In: *International Journal of Medical Informatics* 167, 104878.

Kayser, M. (2012). *Editors' pick: Christmas is coming-time for chocolate to get ready for your Nobel Prize*.

Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. (2017). "Variable importance using decision trees". In: *Advances in neural information processing systems* 30.

Kelley, H. H. (1973). "The processes of causal attribution." In: *American psychologist* 28.2, 107.

Kendall, M. G. (1948). *Rank correlation methods*. Griffin.

Kennedy, A. D. and DeLeo, F. R. (2009). "Neutrophil apoptosis and the resolution of infection". In: *Immunologic research* 43, 25–61.

Khameneh, H. J., Ho, A. W., Laudisi, F., Derks, H., Kandasamy, M., Sivasankar, B., Teng, G. G., and Mortellaro, A. (2017). "C5a regulates IL-1β production and leukocyte recruitment in a murine model of monosodium urate crystal-induced peritonitis". In: *Frontiers in pharmacology* 8, 10.

Khamnuan, P., Chongruksut, W., Jearwattanakanok, K., Patumanond, J., and Tantraworasin, A. (2015). "Necrotizing fasciitis: epidemiology and clinical predictors for amputation". In: *International journal of general medicine*, 195–202.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001). "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks". In: *Nature medicine* 7.6, 673.

Kim, K.-T., Kim, Y. J., Won Lee, J., Kim, Y. J., Park, S.-W., Lim, M. K., and Suh, C. H. (2011). "Can necrotizing infectious fasciitis be differentiated from nonnecrotizing infectious fasciitis with MR imaging?" In: *Radiology* 259.3, 816–824.

Kinasewitz, G. T., Yan, S. B., Basson, B., Comp, P., Russell, J. A., Cariou, A., Um, S. L., Utterback, B., Laterre, P.-F., and Dhainaut, J.-F. (2004). "Universal changes in biomarkers of coagulation and inflammation occur in patients with severe sepsis, regardless of causative micro-organism [ISRCTN74215569]". In: *Critical care* 8.2, 1–9.

King, D., Karthikesalingam, A., Hughes, C., Montgomery, H., Raine, R., Rees, G., and Team, D. H. (2018). "Letter in response to Google DeepMind and healthcare in an age of algorithms". In: *Health and Technology* 8, 11–13.

Kingma, E. (2010). "Paracetamol, poison, and polio: why Boorse's account of function fails to distinguish health and disease". In: *The British Journal for the Philosophy of Science*.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space". In: *Database* 2011.

Kirby, R. M. and Meyer, M. (2013). "Visualization collaborations: What works and why". In: *IEEE computer graphics and applications* 33.6, 82–88.

Kitano, H. (2002). "Systems biology: a brief overview". In: *science* 295.5560, 1662–1664.

Klamt, S., Stelling, J., Ginkel, M., and Gilles, E. D. (2003). "FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps". In: *Bioinformatics* 19.2, 261–269.

Klanderman, R. B., Bosboom, J. J., Migdady, Y., Veelo, D. P., Geerts, B. F., Murphy, M. F., and Vlaar, A. P. (2019). "Transfusion-associated circulatory overload—a systematic review of diagnostic biomarkers". In: *Transfusion* 59.2, 795–805.

Klein, G. A. (2017). *Sources of power: How people make decisions*. MIT press.

Kobayashi, S. D., Braughton, K. R., Whitney, A. R., Voyich, J. M., Schwan, T. G., Musser, J. M., and DeLeo, F. R. (2003). "Bacterial pathogens modulate an apoptosis differentiation program in human neutrophils". In: *Proceedings of the National Academy of Sciences* 100.19, 10948–10953.

Kolaczkowska, E. and Kubes, P. (2013). "Neutrophil recruitment and function in health and inflammation". In: *Nature reviews immunology* 13.3, 159–175.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Korcsmaros, T., Schneider, M. V., and Superti-Furga, G. (2017). "Next generation of network medicine: interdisciplinary signaling approaches". In: *Integrative Biology* 9.2, 97–108.

Korpi, J. T., Åström, P., Lehtonen, N., Tjäderhane, L., Kallio-Pulkkinen, S., Siponen, M., Sorsa, T., Pirilä, E., and Salo, T. (2009). "Healing of extraction sockets in collagenase-2 (matrix metalloproteinase-8)-deficient mice". In: *European journal of oral sciences* 117.3, 248–254.

Koussih, L., Atoui, S., Tliba, O., and Gounni, A. S. (2021). "New Insights on the Role of pentraxin-3 in Allergic Asthma". In: *Frontiers in Allergy* 2, 678023.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). "Estimating mutual information". In: *Phys. Rev. E* 69 (6), 066138. DOI: 10.1103/PhysRevE.69.066138. URL: https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Kristensen, M. K., Hansen, M. B., Madsen, M. B., Hansen, C. B., Pilely, K., Hyldegaard, O., and Garred, P. (2020). "Complement activation is associated with mortality in patients with necrotizing soft-tissue infections—a prospective observational study". In: *Frontiers in immunology* 11, 17.

Kube, T., Blease, C., Ballou, S. K., and Kaptchuk, T. J. (2019). "Hope in medicine: applying multidisciplinary insights". In: *Perspectives in Biology and Medicine* 62.4, 591–616.

Kudo, D., Goto, T., Uchimido, R., Hayakawa, M., Yamakawa, K., Abe, T., Shiraishi, A., and Kushimoto, S. (2021). "Coagulation phenotypes in sepsis and effects of recombinant human thrombomodulin: an analysis of three multicentre observational studies". In: *Critical Care* 25, 1–11.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. C., et al. (2020). "Package 'caret'". In: *The R Journal* 223.7.

Kuijjer, M. L., Hsieh, P.-H., Quackenbush, J., and Glass, K. (2019). "lionessR: single sample network inference in R". In: *BMC cancer* 19.1, 1003.

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. (2019). "Estimating sample-specific regulatory networks". In: *iScience* 14, 226–240.

Kuivaniemi, H. and Tromp, G. (2019). "Type III collagen (COL3A1): Gene and protein structure, tissue distribution, and associated diseases". In: *Gene* 707, 151–171.

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, 79–86.

Kunapuli, G., Varghese, B. A., Ganapathy, P., Desai, B., Cen, S., Aron, M., Gill, I., and Duddalwar, V. (2018). "A decision-support tool for renal mass classification". In: *Journal of Digital Imaging* 31, 929–939.

Kwakkel, J. H. and Pruyt, E. (2013). "Exploratory Modeling and Analysis, an approach for model-based foresight under deep uncertainty". In: *Technological Forecasting and Social Change* 80.3, 419–431.

Kwon, O.-H., Crnovrsanin, T., and Ma, K.-L. (2017). "What would a graph look like in this layout? a machine learning approach to large graph visualization". In: *IEEE transactions on visualization and computer graphics* 24.1, 478–488.

Kyrou, I. and Tsigos, C. (2009). "Stress hormones: physiological stress and regulation of metabolism". In: *Current opinion in pharmacology* 9.6, 787–793.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., et al. (2008). "The virophage as a unique parasite of the giant mimivirus". In: *Nature* 455.7209, 100–104.

Laronha, H. and Caldeira, J. (2020). "Structure and function of human matrix metalloproteinases". In: *Cells* 9.5, 1076.

Lasenby, J., Lasenby, A. N., and Doran, C. J. (2000). "A unified mathematical language for physics and engineering in the 21st century". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358.1765, 21–39.

LaValley, M. P. (2008). "Logistic regression". In: *Circulation* 117.18, 2395–2399.

Lawrence, M. G., Williams, S., Nanz, P., and Renn, O. (2022). "Characteristics, potentials, and challenges of transdisciplinary research". In: *One Earth* 5.1, 44–61.

Le Breton, Y., Belew, A. T., Freiberg, J. A., Sundar, G. S., Islam, E., Lieberman, J., Shirtliff, M. E., Tettelin, H., El-Sayed, N. M., and McIver, K. S. (2017). "Genome-wide discovery of novel M1T1 group A streptococcal determinants important for fitness and virulence during soft-tissue infection". In: *PLoS Pathogens* 13.8, e1006584.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study". In: *Jama* 270.24, 2957–2963.

Lee, J. and Hastie, T. (2013). "Structure learning of mixed graphical models". In: *Artificial Intelligence and Statistics*. PMLR, 388–396.

Lee, S. M. and An, W. S. (2016). "New clinical criteria for septic shock: serum lactate level as new emerging vital sign". In: *Journal of thoracic disease* 8.7, 1388.

Lee, Y. S., Chang, J. Y., and Choi, J. N. (2017). "Why reject creative ideas? Fear as a driver of implicit bias against creativity". In: *Creativity Research Journal* 29.3, 225–235.

Leichtle, S. W., Tung, L., Khan, M., Inaba, K., and Demetriades, D. (2016). "The role of radiologic evaluation in necrotizing soft tissue infections". In: *Journal of Trauma and Acute Care Surgery* 81.5, 921–924.

Leikin, R., Berman, A., and Zaslavsky, O. (2000). "Applications of symmetry to problem solving". In: *International Journal of Mathematical Education in Science and Technology* 31.6, 799–809.

Leist, A. K., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, S. P., Muniz-Terrera, G., and Wade, S. (2022). "Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences". In: *Science Advances* 8.42, eabk1942.

Lerche, V., Neubauer, A. B., and Voss, A. (2018). "Effects of implicit fear of failure on cognitive processing: A diffusion model analysis". In: *Motivation and Emotion* 42, 386–402.
Leung, H. and Haykin, S. (1991). "The complex backpropagation algorithm". In: *IEEE Transactions on signal processing* 39.9, 2101–2104.
Levi, M. and Van Der Poll, T. (2012). "Thrombomodulin in sepsis." In: *Minerva anestesiologica* 79.3, 294–298.
Levi, M. and Poll, T. van der (2008). "The role of natural anticoagulants in the pathogenesis and management of systemic activation of coagulation and inflammation in critically ill patients". In: *Seminars in thrombosis and hemostasis*. Vol. 34. 05. © Thieme Medical Publishers, 459–468.
— (2017). "Coagulation and sepsis". In: *Thrombosis research* 149, 38–44.
Lewandowski, D., Kurowicka, D., and Joe, H. (2009). "Generating random correlation matrices based on vines and extended onion method". In: *Journal of multivariate analysis* 100.9, 1989–2001.
Liang, S. C., Tan, X.-Y., Luxenberg, D. P., Karim, R., Dunussi-Joannopoulos, K., Collins, M., and Fouser, L. A. (2006). "Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides". In: *The Journal of experimental medicine* 203.10, 2271–2279.
Liaw, A., Wiener, M., et al. (2002). "Classification and regression by randomForest". In: *R news* 2.3, 18–22.
Lin, J.-J., Hsiao, H.-J., Chan, O.-W., Wang, Y., Hsia, S.-H., and Chiu, C.-H. (2017). "Increased serum thrombomodulin level is associated with disease severity and mortality in pediatric sepsis". In: *PLoS One* 12.8, e0182324.
Lindfors, E., Dam, J. C. van, Lam, C. M. C., Zondervan, N. A., Martins dos Santos, V. A., and Suarez-Diez, M. (2018). "SyNDI: synchronous network data integration framework". In: *BMC bioinformatics* 19, 1–15.
Lindlöf, A. and Lubovac, Z. (2005). "Simulations of simple artificial genetic networks reveal features in the use of Relevance Networks". In: *In silico biology* 5.3, 239–249.
Ling, X.-W., Zhang, T.-T., Ling, M.-M., Chen, W.-H., Huang, C.-H., and Shen, G.-L. (2023). "Th1/Th2 cytokine levels: A potential diagnostic tool for patients with necrotizing fasciitis". In: *Burns* 49.1, 200–208.
Linn, R. L. and Werts, C. E. (1969). "Assumptions in making causal inferences from part correlations, partal correlations, and partial regression coefficients." In: *Psychological Bulletin* 72.5, 307.
Linnainmaa, S. (1976). "Taylor expansion of the accumulated rounding error". In: *BIT Numerical Mathematics* 16.2, 146–160.
Linthwaite, S. and Fuller, G. N. (2013). "Milk, chocolate and Nobel prizes". In: *Practical neurology* 13.1, 63–63.
Liu, C., Suo, S., Luo, L., Chen, X., Ling, C., and Cao, S. (2022). "SOFA score in relation to sepsis: clinical implications in diagnosis, treatment, and prognostic assessment". In: *Computational and Mathematical Methods in Medicine* 2022.
Liu, R., Wang, X., Aihara, K., and Chen, L. (2014). "Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers". In: *Medicinal research reviews* 34.3, 455–478.

Liu, X., Liu, Z.-P., Zhao, X.-M., and Chen, L. (2012). "Identifying disease genes and module biomarkers by differential interactions". In: *Journal of the American Medical Informatics Association* 19.2, 241–248.

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). "Personalized characterization of diseases using sample-specific networks". In: *Nucleic acids research* 44.22, e164–e164.

Loh, W.-L. (1996). "On Latin hypercube sampling". In: *The annals of statistics* 24.5, 2058–2080.

Loh, W.-Y. (2011). "Classification and regression trees". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.1, 14–23.

Loughman, J. A. and Caparon, M. G. (2006). "A novel adaptation of aldolase regulates virulence in Streptococcus pyogenes". In: *The EMBO journal* 25.22, 5414–5422.

Low, D. E. (2013). "Toxic shock syndrome: major advances in pathogenesis, but not treatment". In: *Critical care clinics* 29.3, 651–675.

Lu, H. and Wang, M. (2019). "RL4health: crowdsourcing reinforcement learning for knee replacement pathway optimization". In: *arXiv preprint arXiv:1906.01407*.

Lu, X.-L., Cai, J.-T., Lu, X.-G., Si, J.-M., and Qian, K.-D. (2007). "Plasma level of thrombomodulin is an early indication of pancreatic necrosis in patients with acute pancreatitis". In: *Internal Medicine* 46.8, 441–446.

Luchian, I., Goriuc, A., Sandu, D., and Covasa, M. (2022). "The role of matrix metalloproteinases (MMP-8, MMP-9, MMP-13) in periodontal and peri-implant pathological processes". In: *International Journal of Molecular Sciences* 23.3, 1806.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). "Consistent individualized feature attribution for tree ensembles". In: *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Lungstras-Bufler, K., Bufler, P., Abdullah, R., Rutherford, C., Endres, S., Abraham, E., Dinarello, C. A., and Rodriguez, R. M. (2004). "High cytokine levels at admission are associated with fatal outcome in patients with necrotizing fasciitis". In: *European cytokine network* 15.2, 135–138.

Lusczek, E. R., Lexcen, D. R., Witowski, N. E., Mulier, K. E., and Beilman, G. (2013). "Urinary metabolic network analysis in trauma, hemorrhagic shock, and resuscitation". In: *Metabolomics* 9.1, 223–235.

Lynskey, N. N., Reglinski, M., Calay, D., Siggins, M. K., Mason, J. C., Botto, M., and Sriskandan, S. (2017). "Multi-functional mechanisms of immune evasion by the streptococcal complement inhibitor C5a peptidase". In: *PLoS pathogens* 13.8, e1006493.

Ma'ayan, A. (2011). "Introduction to network analysis in systems biology." In: *Science signaling* 4.190, tr5. DOI: 10.1126/scisignal.2001965. URL: http://www.ncbi.nlm.nih.gov/pubmed/21917719%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3196357.

Maćkiewicz, A. and Ratajczak, W. (1993). "Principal components analysis (PCA)". In: *Computers & Geosciences* 19.3, 303–342.

Mackman, N. (2009). "The many faces of tissue factor". In: *Journal of Thrombosis and Haemostasis* 7, 136–139.

MacLeod, M. (2018). "What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice". In: *Synthese* 195.2, 697–720.

Madsen, M. B., Hjortrup, P. B., Hansen, M. B., Lange, T., Norrby-Teglund, A., Hyldegaard, O., and Perner, A. (2017). "Immunoglobulin G for patients with necrotising soft tissue infection (INSTINCT): a randomised, blinded, placebo-controlled trial". In: *Intensive care medicine* 43, 1585–1593.

Madsen, M. B., Arnell, P., and Hyldegaard, O. (2020). "Necrotizing Soft-Tissue Infections: Clinical Features and Diagnostic Aspects". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 39–52.

Madsen, M. B., Bergsten, H., and Norrby-Teglund, A. (2020). "Treatment of necrotizing soft tissue infections: IVIG". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 105–125.

Madsen, M. B., Skrede, S., Perner, A., Arnell, P., Nekludov, M., Bruun, T., Karlsson, Y., Hansen, M. B., Polzik, P., Hedetoft, M., et al. (2019). "Patient's characteristics and outcomes in necrotising soft-tissue infections: results from a Scandinavian, multicentre, prospective cohort study". In: *Intensive Care Medicine* 45, 1241–1251.

Madsen, M., Skrede, S., Bruun, T., Arnell, P., Rosén, A., Nekludov, M., Karlsson, Y., Bergey, F., Saccenti, E., Martins dos Santos, V., et al. (2018). "Necrotizing soft tissue infections–a multicentre, prospective observational study (INFECT): protocol and statistical analysis plan". In: *Acta Anaesthesiologica Scandinavica* 62.2, 272–279.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I., Yakunin, A. F., et al. (2011). "Evolution and classification of the CRISPR–Cas systems". In: *Nature Reviews Microbiology* 9.6, 467–477.

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V., Tiwari, K., Roberts, M. G., Xavier, A., Vu, M. T., Men, J., Maire, M., Kananathan, S., et al. (2020). "BioModels—15 years of sharing computational models in life science". In: *Nucleic Acids Research* 48.D1, D407–D415.

Manovich, L. (2011). "What is visualisation?" In: *Visual Studies* 26.1, 36–49.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., et al. (2012). "Wisdom of crowds for robust gene network inference". In: *Nature methods* 9.8, 796–804.

Martins dos Santos, V. A., Hardt, C., Skrede, S., and Saccenti, E. (2020). "Systems and precision medicine in necrotizing soft tissue infections". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 187–207.

Marwick, C., Broomhall, J., McCowan, C., Phillips, G., Gonzalez-McQuire, S., Akhras, K., Merchant, S., Nathwani, D., and Davey, P. (2011). "Severity assessment of skin and soft tissue infections: cohort study of management and outcomes for hospitalized patients". In: *Journal of Antimicrobial Chemotherapy* 66.2, 387–397.

Mason, M. J., Fan, G., Plath, K., Zhou, Q., and Horvath, S. (2009). "Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells". In: *BMC genomics* 10.1, 327.

MATLAB (2018). *version 9.5.0 (R2018b)*. Natick, Massachusetts: The MathWorks Inc.

— (2022). *version 9.13 (R2022b)*. Natick, Massachusetts: The MathWorks Inc.

Matsuda, H. (2000). "Physical nature of higher-order mutual information: Intrinsic correlations and frustration". In: *Physical review E* 62.3, 3096.

Matzinger, P. (2002). "The danger model: a renewed sense of self". In: *science* 296.5566, 301–305.

May, A. K. (2009). "Skin and soft tissue infections". In: *Surgical Clinics of North America* 89.2, 403–420.
— (2011). "Skin and soft tissue infections: the new surgical infection society guidelines". In: *Surgical infections* 12.3, 179–184.
May, A. K., Stafford, R. E., Bulger, E. M., Heffernan, D., Guillamondegui, O., Bochicchio, G., and Eachempati, S. R. (2009). "Treatment of complicated skin and soft tissue infections". In: *Surgical infections* 10.5, 467–499.
McCarthy, J. (2004). "Tackling the challenges of interdisciplinary bioscience". In: *Nature Reviews Molecular Cell Biology* 5.11, 933–937.
McHenry, C. R., Piotrowski, J. J., Petrinic, D., and Malangoni, M. A. (1995). "Determinants of mortality for necrotizing soft-tissue infections." In: *Annals of surgery* 221.5, 558.
McIver, K. S., Subbarao, S., Kellner, E. M., Heath, A. S., and Scott, J. R. (1996). "Identification of isp, a locus encoding an immunogenic secreted protein conserved among group A streptococci". In: *Infection and immunity* 64.7, 2548–2555.
McLaughlin, K. (2016). *Empowerment: A critique*. Routledge.
Mcnicholas, P. D. and Murphy, T. B. (2008). "Parsimonious Gaussian mixture models". In: *Statistics and Computing* 18.3, 285–296.
Med, C. I. (2006). "Multidisciplinarity, interdisciplinarity and transdis-ciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness". In: *Clin Invest Med* 29.6, 351–364.
Medina, L. M. P. et al. (2021). "Discriminatory plasma biomarkers predict specific clinical phenotypes of necrotizing soft-tissue infections". In: *Journal of Clinical Investigation* 131 (14). DOI: 10.1172/JCI149523. URL: https://www.jci.org/articles/view/149523.
Medsker, L. R. and Jain, L. (2001). "Recurrent neural networks". In: *Design and Applications* 5, 64–67.
Meinshausen, N. and Bühlmann, P. (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, 417–473.
Meleney, F. L. (1924). "Hemolytic streptococcus gangrene". In: *Archives of Surgery* 9.2, 317–364.
Mendez, K. M., Reinke, S. N., and Broadhurst, D. I. (2019). "A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification". In: *Metabolomics* 15.12, 150.
Merle, N., Church, S., Fremeaux-Bacchi, V., and Roumenina, L. (2015). *Complement system part I-Molecular mechanisms of activation and regulation. Front. Immunol.(2015)*.
Messerli, F. H. (2012). "Chocolate consumption, cognitive function, and Nobel laureates". In: *N Engl J Med* 367.16, 1562–1564.
Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). "The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes". In: *BMC bioinformatics* 9.1, 386.
Meyer, P. E. (2008). "Information-theoretic variable selection and network inference from microarray data". In: *Universite Libre de Bruxelles.[Google Scholar]*.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools". In: *Nucleic acids research* 47.D1, D419–D426.
Miao, H., Xia, X., Perelson, A. S., and Wu, H. (2011). "On identifiability of nonlinear ODE models and applications in viral dynamics". In: *SIAM review* 53.1, 3–39.
Michotte, A. (2017). *The perception of causality*. Vol. 21. Routledge.
Mihajlovic, D. M., Lendak, D. F., Draskovic, B. G., Mikic, A. S. N., Mitic, G. P., Cebovic, T. N., and Brkic, S. V. (2015). "Thrombomodulin is a strong predictor of multiorgan dysfunction syndrome in patients with sepsis". In: *Clinical and Applied Thrombosis/Hemostasis* 21.5, 469–474.
Mihara, M., Hashizume, M., Yoshida, H., Suzuki, M., and Shiina, M. (2012). "IL-6/IL-6 receptor system and its role in physiological and pathological conditions". In: *Clinical science* 122.4, 143–159.
Miller, L. G., Perdreau-Remington, F., Rieg, G., Mehdi, S., Perlroth, J., Bayer, A. S., Tang, A. W., Phung, T. O., and Spellberg, B. (2005). "Necrotizing fasciitis caused by community-associated methicillin-resistant Staphylococcus aureus in Los Angeles". In: *New England Journal of Medicine* 352.14, 1445–1453.
Minai-Fleminger, Y. and Levi-Schaffer, F. (2009). "Mast cells and eosinophils: the two key effector cells in allergic inflammation". In: *Inflammation research* 58, 631–638.
Miranda, D. and Bulger, E. M. (2018). "Novel immune therapies in the management of streptococcal sepsis and necrotizing soft tissue infections". In: *Surgical Infections* 19.8, 745–749.
Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI". In: *Nature machine intelligence* 1.11, 501–507.
Moats, D. (2021). "Rethinking the 'Great Divide': Approaching interdisciplinary collaborations around digital data with humour and irony". In: *Science & Technology Studies* 34.1, 19–42.
Molloy, E. M., Cotter, P. D., Hill, C., Mitchell, D. A., and Ross, R. P. (2011). "Streptolysin S-like virulence factors: the continuing sagA". In: *Nature Reviews Microbiology* 9.9, 670–681.
Monteiro, M. and Keating, E. (2009). "Managing misunderstandings: The role of language in interdisciplinary scientific collaboration". In: *Science communication* 31.1, 6–28.
Moore, K. L., Esmon, C. T., and Esmon, N. L. (1989). "Tumor necrosis factor leads to the internalization and degradation of thrombomodulin from the surface of bovine aortic endothelial cells in culture". In.
Morgan, M. (2010). "Diagnosis and management of necrotising fasciitis: a multiparametric approach". In: *Journal of Hospital Infection* 75.4, 249–257.
Morley, J. and Floridi, L. (2021). "How to design a governable digital health ecosystem". In: *The 2020 Yearbook of the Digital Ethics Lab*, 69–88.
Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., and Floridi, L. (2020). "The ethics of AI in health care: a mapping review". In: *Social Science & Medicine* 260, 113172.
Morris, M. W. and Larrick, R. P. (1995). "When one cause casts doubt on another: A normative analysis of discounting in causal attribution." In: *Psychological Review* 102.2, 331.

Morrison, E. W. (2011). "Employee voice behavior: Integration and directions for future research". In: *Academy of Management annals* 5.1, 373–412.
Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT press.
Murtagh, F. and Contreras, P. (2012). "Algorithms for hierarchical clustering: an overview". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1, 86–97.
Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., and Brown, S. D. (2004). "An introduction to decision tree modeling". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 18.6, 275–285.
Naegeli, A., Bratanis, E., Karlsson, C., Shannon, O., Kalluru, R., Linder, A., Malmström, J., and Collin, M. (2019). "Streptococcus pyogenes evades adaptive immunity through specific IgG glycan hydrolysis". In: *Journal of Experimental Medicine* 216.7, 1615–1629.
Nagata, S. (1999). "Fas ligand-induced apoptosis". In: *Annual review of genetics* 33.1, 29–55.
Nakhla, G. (2019). "The relationship between fear of failure, academic motivation and student engagement in higher education:: A general linear model". PhD thesis. Lancaster University.
Naseer, U., Steinbakk, M., Blystad, H., and Caugant, D. (2016). "Epidemiology of invasive group A streptococcal infections in Norway 2010–2014: a retrospective cohort study". In: *European Journal Of Clinical Microbiology & Infectious Diseases* 35, 1639–1648.
Natekin, A. and Knoll, A. (2013). "Gradient boosting machines, a tutorial". In: *Frontiers in neurorobotics* 7, 21.
Nawijn, F., Smeeing, D. P., Houwert, R. M., Leenen, L. P., and Hietbrink, F. (2020). "Time is of the essence when treating necrotizing soft tissue infections: a systematic review and meta-analysis". In: *World Journal of Emergency Surgery* 15, 1–11.
Nedrebo, T. and Skrede, S. (2020). "Necrotizing Soft Tissue Infections: Case Reports, from the Clinician's Perspectives". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 21–37.
Neeki, M. M., Dong, F., Au, C., Toy, J., Khoshab, N., Lee, C., Kwong, E., Yuen, H. W., Lee, J., Ayvazian, A., et al. (2017). "Evaluating the laboratory risk indicator to differentiate cellulitis from necrotizing fasciitis in the emergency department". In: *Western Journal of Emergency Medicine* 18.4, 684.
Nelson, A., Herron, D., Rees, G., and Nachev, P. (2019). "Predicting scheduled hospital attendance with artificial intelligence". In: *NPJ digital medicine* 2.1, 26.
Nelson, G. E., Pondo, T., Toews, K.-A., Farley, M. M., Lindegren, M. L., Lynfield, R., Aragon, D., Zansky, S. M., Watt, J. P., Cieslak, P. R., et al. (2016). "Epidemiology of invasive group A streptococcal infections in the United States, 2005–2012". In: *Reviews of Infectious Diseases* 63.4, 478–486.
Nemenman, I., Bialek, W., and Van Steveninck, R. D. R. (2004). "Entropy and information in neural spike trains: Progress on the sampling problem". In: *Physical Review E* 69.5, 056111.
Neumann, A., Happonen, L., Karlsson, C., Bahnan, W., Frick, I.-M., and Björck, L. (2021). "Streptococcal protein SIC activates monocytes and induces inflammation". In: *Iscience* 24.4, 102339.

Noble, D. (2008). "Claude Bernard, the first systems biologist, and the future of physiology". In: *Experimental physiology* 93.1, 16–26.
Norrby-Teglund, A., Chatellier, S., Low, D. E., McGeer, A., Green, K., and Kotb, M. (2000). "Host variation in cytokine responses to superantigens determine the severity of invasive group A streptococcal infection". In: *European journal of immunology* 30.11, 3247–3255.
Norrby-Teglund, A., Thulin, P., Gan, B. S., Kotb, M., McGeer, A., Andersson, J., and Low, D. E. (2001). "Evidence for superantigen involvement in severe group a streptococcal tissue infections". In: *The Journal of infectious diseases* 184.7, 853–860.
Numata, J., Ebenhöh, O., and Knapp, E.-W. (2008). "Measuring correlations in metabolomic networks with mutual information". In: *Genome Informatics 2008: Genome Informatics Series Vol. 20*. World Scientific, 112–122.
O'Loughlin, A. and McFadzean, E. (1999). "Toward a holistic theory of strategic problem solving". In: *Team Performance Management: An International Journal* 5.3, 103–120.
Oeberst, A. and Imhoff, R. (2023). "Toward Parsimony in Bias Research: A Proposed Common Framework of Belief-Consistent Information Processing for a Set of Biases". In: *Perspectives on Psychological Science*, 17456916221148147.
Ogle, D., Wheeler, P., and Dinno, A. (2020). *FSA: fisheries stock analysis. R package version 0.8. 30*.
Ogunniyi, A. D., Grabowicz, M., Mahdi, L. K., Cook, J., Gordon, D. L., Sadlon, T. A., and Paton, J. C. (2009). "Pneumococcal histidine triad proteins are regulated by the Zn2+-dependent repressor AdcR and inhibit complement deposition through the recruitment of complement factor H". In: *The FASEB Journal* 23.3, 731–738.
Okamoto, T., Tanigami, H., Suzuki, K., Shimaoka, M., et al. (2012). "Thrombomodulin: a bifunctional modulator of inflammation and coagulation in sepsis". In: *Critical care research and practice* 2012.
Okasha, S. (2002). *Philosophy of science: A very short introduction*. Vol. 67. Oxford Paperbacks.
Opgen-Rhein, R. and Strimmer, K. (2006). "Inferring gene dependency networks from genomic longitudinal data: a functional data approach". In: *REVSTAT-Statistical Journal* 4.1, 53–65.
— (2007). "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data". In: *BMC systems biology* 1.1, 1–10.
Oppegaard, O., Mylvaganam, H., and Kittang, B. (2015). "Beta-haemolytic group A, C and G streptococcal infections in Western Norway: a 15-year retrospective survey". In: *Clinical Microbiology and Infection* 21.2, 171–178.
Osborne, J. (2010). "Arguing to learn in science: The role of collaborative, critical discourse". In: *science* 328.5977, 463–466.
Osowicki, J., Azzopardi, K. I., Fabri, L., Frost, H. R., Rivera-Hernandez, T., Neeland, M. R., Whitcombe, A. L., Grobler, A., Gutman, S. J., Baker, C., et al. (2021). "A controlled human infection model of Streptococcus pyogenes pharyngitis (CHIVAS-M75): an observational, dose-finding study". In: *The Lancet Microbe* 2.7, e291–e299.

Oud, L. and Watkins, P. (2015). "Contemporary trends of the epidemiology, clinical characteristics, and resource utilization of necrotizing fasciitis in Texas: a population-based cohort study". In: *Critical care research and practice* 2015.

Oxenham, M. F., Tilley, L., Matsumura, H., Nguyen, L. C., Nguyen, K. T., Nguyen, K. D., Domett, K., and Huffer, D. (2009). "Paralysis and severe disability requiring intensive care in Neolithic Asia". In: *Anthropological Science* 117.2, 107–112.

Ozyaprak, M. (2016). "The effectiveness of SCAMPER technique on creative thinking skills". In: *Journal for the Education of Gifted young scientists* 4.1, 31–40.

Paahlman, L. I., Malmström, E., Mörgelin, M., and Herwald, H. (2007). "M protein from Streptococcus pyogenes induces tissue factor expression and pro-coagulant activity in human monocytes". In: *Microbiology* 153.Pt 8, 2458.

Page, S. E. (2019). *The diversity bonus: How great teams pay off in the knowledge economy*. Princeton University Press.

Palm, F., Chowdhury, S., Wettemark, S., Malmström, J., Happonen, L., and Shannon, O. (2022). "Distinct serotypes of streptococcal M proteins mediate fibrinogen-dependent platelet activation and proinflammatory effects". In: *Infection and Immunity* 90.2, e00462–21.

Pamp, S. J., Frees, D., Engelmann, S., Hecker, M., and Ingmer, H. (2006). "Spx is a global effector impacting stress tolerance and biofilm formation in Staphylococcus aureus". In: *Journal of bacteriology* 188.13, 4861–4870.

Pancholi, V. and Caparon, M. (2022). "Streptococcus pyogenes metabolism". In.

Paninski, L. (2003). "Estimation of entropy and mutual information". In: *Neural computation* 15.6, 1191–1253.

Pannell, J., Dencer-Brown, A., Greening, S., Hume, E., Jarvis, R., Mathieu, C., Mugford, J., and Runghen, R. (2019). "An early career perspective on encouraging collaborative and interdisciplinary research in ecology". In: *Ecosphere* 10.10, e02899.

Pannucci, C. J. and Wilkins, E. G. (2010). "Identifying and avoiding bias in research". In: *Plastic and reconstructive surgery* 126.2, 619.

Pasko, L. (2010). "Damaged daughters: The history of girls' sexuality and the juvenile justice system". In: *The Journal of Criminal Law and Criminology*, 1099–1130.

Paul, R., Ariey, F., and Robert, V. (2003). "The evolutionary ecology of Plasmodium". In: *Ecology Letters* 6.9, 866–880.

Pauli, J. N., Mendoza, J. E., Steffan, S. A., Carey, C. C., Weimer, P. J., and Peery, M. Z. (2014). "A syndrome of mutualism reinforces the lifestyle of a sloth". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1778, 20133006.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Pearson, K. (1895). "VII. Note on regression and inheritance in the case of two parents". In: *proceedings of the royal society of London* 58.347-352, 240–242.

Peetermans, M., Prost, N. de, Eckmann, C., Norrby-Teglund, A., Skrede, S., and De Waele, J. (2020). "Necrotizing skin and soft-tissue infections in the intensive care unit". In: *Clinical Microbiology and Infection* 26.1, 8–17.

Pennington, J., Schoenholz, S., and Ganguli, S. (2017). "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice". In: *Advances in neural information processing systems* 30.

Pennington, N. and Hastie, R. (1986). "Evidence evaluation in complex decision making." In: *Journal of personality and social psychology* 51.2, 242.

— (1992). "Explaining the evidence: Tests of the Story Model for juror decision making." In: *Journal of personality and social psychology* 62.2, 189.
Petersen, A.-K., Krumsiek, J., Wägele, B., Theis, F. J., Wichmann, H.-E., Gieger, C., and Suhre, K. (2012). "On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies". In: *BMC bioinformatics* 13, 1–7.
Pham, T. N., Moore, M. L., Costa, B. A., Cuschieri, J., and Klein, M. B. (2009). "Assessment of functional limitation after necrotizing soft tissue infection". In: *Journal of burn care & research* 30.2, 301–306.
Pierrakos, C., Velissaris, D., Bisdorff, M., Marshall, J. C., and Vincent, J.-L. (2020). "Biomarkers of sepsis: time for a reappraisal". In: *Critical Care* 24.1, 1–15.
Plainvert, C., Doloy, A., Loubinoux, J., Lepoutre, A., Collobert, G., Touak, G., Trieu-Cuot, P., Bouvet, A., and Poyart, C. (2012). "Invasive group A streptococcal infections in adults, France (2006–2010)". In: *Clinical microbiology and infection* 18.7, 702–710.
Ploug, T. and Holm, S. (2020). "The right to refuse diagnostics and treatment planning by artificial intelligence". In: *Medicine, Health Care and Philosophy* 23.1, 107–114.
Polzik, P., Grøndal, O., Tavenier, J., Madsen, M. B., Andersen, O., Hedetoft, M., and Hyldegaard, O. (2019). "SuPAR correlates with mortality and clinical severity in patients with necrotizing soft-tissue infections: results from a prospective, observational cohort study". In: *Scientific Reports* 9.1, 1–8.
Poulin, R. and Morand, S. (2000). "The diversity of parasites". In: *The quarterly review of biology* 75.3, 277–293.
Powell, R. and Scarffe, E. (2019). "Rethinking "Disease": a fresh diagnosis and a new philosophical treatment". In: *Journal of Medical Ethics* 45.9, 579–588.
Powers, R. K., Culp-Hill, R., Ludwig, M. P., Smith, K. P., Waugh, K. A., Minter, R., Tuttle, K. D., Lewis, H. C., Rachubinski, A. L., Granrath, R. E., et al. (2019). "Trisomy 21 activates the kynurenine pathway via increased dosage of interferon receptors". In: *Nature communications* 10.1, 1–11.
Powles, J. and Hodson, H. (2017). "Google DeepMind and healthcare in an age of algorithms". In: *Health and technology* 7.4, 351–367.
— (2018). "Response to deepmind". In: *Health and Technology* 8.1-2, 15–29.
Preissner, K. T., Delvos, U., and Mueller-Berghaus, G. (1987). "Binding of thrombin to thrombomodulin accelerates inhibition of the enzyme by antithrombin III. Evidence for a heparin-independent mechanism". In: *Biochemistry* 26.9, 2521–2528.
Presslee, S., Slater, G. J., Pujos, F., Forasiepi, A. M., Fischer, R., Molloy, K., Mackie, M., Olsen, J. V., Kramarz, A., Taglioretti, M., et al. (2019). "Palaeoproteomics resolves sloth relationships". In: *Nature Ecology & Evolution* 3.7, 1121–1130.
Proft, T. and Fraser, J. D. (2022). "Streptococcal superantigens: biological properties and potential role in disease". In: *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet]. 2nd edition*.
Purchase, H. C. (2014). "Twelve years of diagrams research". In: *Journal of Visual Languages & Computing* 25.2, 57–75.
Putnam, L. R., Richards, M. K., Sandvall, B. K., Hopper, R. A., Waldhausen, J. H., and Harting, M. T. (2016). "Laboratory evaluation for pediatric patients with

suspected necrotizing soft tissue infections: A case–control study". In: *Journal of Pediatric Surgery* 51.6, 1022–1025.

Python Core Team (2015). "Python: A dynamic, open source programming language." In: *Python Software Foundation*. 78. URL: www.python.org.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: www.R-project.org.

Racine, E., Boehlen, W., and Sample, M. (2019). "Healthcare uses of artificial intelligence: Challenges and opportunities for growth". In: *Healthcare management forum*. Vol. 32. 5. SAGE Publications Sage CA: Los Angeles, CA, 272–275.

Rahmouni, A., Chosidow, O., Mathieu, D., Gueorguieva, E., Jazaerli, N., Radier, C., Faivre, J., Roujeau, J., and Vasile, N. (1994). "MR imaging in acute infectious cellulitis." In: *Radiology* 192.2, 493–496.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941*.

Rasamoelina, A. D., Adjailia, F., and Sinčák, P. (2020). "A review of activation function for artificial neural network". In: *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 281–286.

Rath, E., Medina, L. M. P., Jahagirdar, S., Mosevoll, K. A., Damås, J. K., Madsen, M. B., Svensson, M., Hyldegaard, O., Dos Santos, V. A. M., Saccenti, E., et al. (2023). "Systemic immune activation profiles in streptococcal necrotizing soft tissue infections: A prospective multicenter study". In: *Clinical Immunology*, 109276.

Ray, B., Ghedin, E., and Chunara, R. (2016). "Network inference from multimodal data: a review of approaches from infectious disease transmission". In: *Journal of biomedical informatics* 64, 44–54.

Reed, S. K. (2016). "The structure of ill-structured (and well-structured) problems revisited". In: *Educational Psychology Review* 28, 691–716.

Rees, J. (1949). "Discussion: Recent techniques in psychosurgery [First session]". In: *Proceedings of the Royal Society of Medicine* 42, 13–22.

Remmington, A. and Turner, C. E. (2018). "The DNases of pathogenic Lancefield streptococci". In: *Microbiology* 164.3, 242–250.

Rhue, L. (2018). "Racial influence on automated perceptions of emotions". In: *Available at SSRN 3281765*.

Ribardo, D. A., Lambert, T. J., and McIver, K. S. (2004). "Role of Streptococcus pyogenes two-component response regulators in the temporal control of Mga and the Mga-regulated virulence gene emm". In: *Infection and immunity* 72.6, 3668–3673.

Richter, J., Brouwer, S., Schroder, K., and Walker, M. J. (2021). "Inflammasome activation and IL-1β signalling in group A Streptococcus disease". In: *Cellular Microbiology* 23.9, e13373.

Riedemann, N. C., Guo, R.-F., Hollmann, T. J., Gao, H., Neff, T. A., Reuben, J. S., Speyer, C. L., Sarma, J. V., Wetsel, R. A., Zetoune, F. S., et al. (2004). "Regulatory role of C5a in LPS-induced IL-6 production by neutrophils during sepsis". In: *The FASEB journal* 18.2, 1–16.

Rijn, J. van (2016). "Massively collaborative machine learning". PhD thesis. Leiden University.

Rist, M. J., Roth, A., Frommherz, L., Weinert, C. H., Krüger, R., Merz, B., Bunzel, D., Mack, C., Egert, B., Bub, A., et al. (2017). "Metabolite patterns predicting sex

and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study". In: *PloS one* 12.8, e0183228.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC bioinformatics* 12.1, 1–8.
Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLoS computational biology* 13.11, e1005752.
Rokach, L. and Maimon, O. (2005). *Clustering methods*.
Romer, A. S. (1967). "Major steps in vertebrate evolution". In: *Science* 158.3809, 1629–1637.
Romero, E. J. and Cruthirds, K. W. (2006). "The use of humor in the workplace". In: *Academy of management perspectives* 20.2, 58–69.
Rosato, A., Tenori, L., Cascante, M., Carulla, P. R. D. A., Santos, V. A. M. dos, and Saccenti, E. (2018). "From correlation to causation: analysis of metabolomics data using systems biology approaches". In: *Metabolomics* 14.4, 37.
Rosato, A., Tenori, L., Cascante, M., De Atauri Carulla, P. R., Martins dos Santos, V. A., and Saccenti, E. (2018). "From correlation to causation: analysis of metabolomics data using systems biology approaches". In: *Metabolomics* 14, 1–20.
Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., Tanguay-Sela, M., Mehltretter, J., Fratila, R., Israel, S., et al. (2021). "Big Data analytics and artificial intelligence in mental healthcare". In: *Applications of Big Data in Healthcare*. Elsevier, 137–171.
Rosmalen, R. P. van, Martins dos Santos, V. A. P., and Suarez-Diez, M. (2022). "Questions, data and models underpinning metabolic engineering". In: *Frontiers in Systems Biology* 2. DOI: 10.3389/fsysb.2022.998048. URL: https://www.frontiersin.org/articles/10.3389/fsysb.2022.998048.
Rosner, K., Ross, C., Karlsmark, T., and Skovgaard, G. L. (2001). "Role of LFA-1/ICAM-1, CLA/E-selectin and VLA-4/VCAM-1 pathways in recruiting leukocytes to the various regions of the chronic leg ulcer". In: *Acta dermato-venereologica* 81.5, 334–339.
Ross, L. N. (2021). "Causal concepts in biology: How pathways differ from mechanisms and why it matters". In: *The British Journal for the Philosophy of Science*.
Ruckenstein, M. and Schüll, N. D. (2017). "The datafication of health". In: *Annual review of anthropology* 46, 261–278.
Rueppell, O., Hayworth, M. K., and Ross, N. (2010). "Altruistic self-removal of health-compromised honey bee workers from their hive". In: *Journal of evolutionary biology* 23.7, 1538–1546.
Runco, M. A. and Acar, S. (2012). "Divergent thinking as an indicator of creative potential". In: *Creativity research journal* 24.1, 66–75.
Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Munoz-Mari, J., et al. (2019). "Inferring causation from time series in Earth system sciences". In: *Nature communications* 10.1, 2553.
Russell, C. D., Parajuli, A., Gale, H. J., Bulteel, N. S., Schuetz, P., Jager, C. P. de, Loonen, A. J., Merekoulias, G. I., and Baillie, J. K. (2019). "The utility of peripheral blood

leucocyte ratios as biomarkers in infectious diseases: A systematic review and meta-analysis". In: *Journal of Infection* 78.5, 339–348.
Saccenti, E. (2017). "Correlation patterns in experimental data are affected by normalization procedures: consequences for data analysis and network inference". In: *Journal of Proteome Research* 16.2, 619–634.
Saccenti, E., Suarez-Diez, M., Luchinat, C., Santucci, C., and Tenori, L. (2015). "Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk". In: *Journal of proteome research* 14.2, 1101–1111.
Saccenti, E. and Svensson, M. (2020). "Systems biology and biomarkers in necrotizing soft tissue infections". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 167–186.
Sakanaka, A., Kuboniwa, M., Hashino, E., Bamba, T., Fukusaki, E., and Amano, A. (2017). "Distinct signatures of dental plaque metabolic byproducts dictated by periodontal inflammatory status". In: *Scientific reports* 7, 42818.
Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). "Explaining deep neural networks and beyond: A review of methods and applications". In: *Proceedings of the IEEE* 109.3, 247–278.
Samuel, A. L. (1959). "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3, 210–229.
Santos, F., Nequiz, M., Hernández-Cuevas, N. A., Hernandez, K., Pineda, E., Encalada, R., Guillen, N., Luis-Garcia, E., Saralegui, A., Saavedra, E., et al. (2015). "Maintenance of intracellular hypoxia and adequate heat shock response are essential requirements for pathogenicity and virulence of Entamoeba histolytica". In: *Cellular microbiology* 17.7, 1037–1051.
Sartelli, M., Guirao, X., Hardcastle, T. C., Kluger, Y., Boermeester, M., Raşa, K., Ansaloni, L., Coccolini, F., Montravers, P., Abu-Zidan, F. M., et al. (2018). "2018 WSES/SIS-E consensus conference: recommendations for the management of skin and soft-tissue infections". In: *World Journal of Emergency Surgery* 13.1, 1–24.
Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006). "Reverse engineering genetic networks using the GeneNet package". In: *The Newsletter of the R Project Volume 6/5, December 2006* 6.9, 50.
Schäfer, J. and Strimmer, K. (2005a). "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Statistical applications in genetics and molecular biology* 4.1.
— (2005b). "An empirical Bayes approach to inferring large-scale gene association networks". In: *Bioinformatics* 21.6, 754–764.
Schmidt, S. R. (2002). "The humour effect: Differential processing and privileged retrieval". In: *Memory* 10.2, 127–138.
Schmidt, S. R. and Williams, A. R. (2001). "Memory for humorous cartoons". In: *Memory & Cognition* 29, 305–311.
Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
Schroeder, J. T. (2009). "Basophils: beyond effector cells of allergic inflammation". In: *Advances in immunology* 101, 123–161.
Schürmann, T. and Grassberger, P. (1996). "Entropy estimation of symbol sequences". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 6.3, 414–427.

Seder, R. A., Paul, W. E., Davis, M. M., and Fazekas de St Groth, B. (1992). "The presence of interleukin 4 during in vitro priming determines the lymphokine-producing potential of CD4+ T cells from T cell receptor transgenic mice." In: *The Journal of experimental medicine* 176.4, 1091–1098.

Sedgewick, A. J., Buschur, K., Shi, I., Ramsey, J. D., Raghu, V. K., Manatakis, D. V., Zhang, Y., Bon, J., Chandra, D., Karoleski, C., et al. (2019). "Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis". In: *Bioinformatics* 35.7, 1204–1212.

Seidl, K., Goerke, C., Wolz, C., Mack, D., Berger-Bächi, B., and Bischoff, M. (2008). "Staphylococcus aureus CcpA affects biofilm formation". In: *Infection and immunity* 76.5, 2044–2050.

Serrat, O. and Serrat, O. (2017). "The SCAMPER technique". In: *Knowledge solutions: tools, methods, and approaches to drive organizational performance*, 311–314.

Setty, S., Cramwinckel, M. J., Nes, E. H. van, Leemput, I. A. van de, Dijkstra, H. A., Lourens, L. J., Scheffer, M., and Sluijs, A. (2023). "Loss of Earth system resilience during early Eocene transient global warming events". In: *Science advances* 9.14, eade5466.

Sfard, A. (2007). "When the rules of discourse change, but nobody tells you: Making sense of mathematics learning from a commognitive standpoint". In: *The journal of the learning sciences* 16.4, 565–613.

Sforzini, L., Nettis, M. A., Mondelli, V., and Pariante, C. M. (2019). "Inflammation in cancer and depression: a starring role for the kynurenine pathway". In: *Psychopharmacology* 236, 2997–3011.

Shampine, L. F. and Reichelt, M. W. (1997). "The matlab ode suite". In: *SIAM journal on scientific computing* 18.1, 1–22.

Shankar-Hari, M., Spencer, J., Sewell, W. A., Rowan, K. M., and Singer, M. (2012). "Bench-to-bedside review: Immunoglobulin therapy for sepsis-biological plausibility from a critical care perspective". In: *Critical care* 16.2, 1–14.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome research* 13.11, 2498–2504.

Sharma, O., O'Seaghdha, M., Velarde, J. J., and Wessels, M. R. (2016). "NAD+-glycohydrolase promotes intracellular survival of group A Streptococcus". In: *PLoS pathogens* 12.3, e1005468.

Sharp, G. C., Ma, H., Saunders, P. T. K., and Norman, J. E. (2013). "A Computational Model of Lipopolysaccharide-Induced Nuclear Factor Kappa B Activation: A Key Signalling Pathway in Infection-Induced Preterm Labour". In: *PLOS ONE* 8.7, 1–7. DOI: 10.1371/journal.pone.0070180. URL: https://doi.org/10.1371/journal.pone.0070180.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli". In: *Nature genetics* 31.1, 64–68.

Shumba, P., Mairpady Shambat, S., and Siemens, N. (2019). "The role of streptococcal and staphylococcal exotoxins and proteases in human necrotizing soft tissue infections". In: *Toxins* 11.6, 332.

Sidiropoulos, K., Viteri, G., Sevilla, C., Jupe, S., Webber, M., Orlic-Milacic, M., Jassal, B., May, B., Shamovsky, V., Duenas, C., et al. (2017). "Reactome enhanced pathway visualization". In: *Bioinformatics* 33.21, 3461–3467.

Siemens, N., Chakrakodi, B., Shambat, S. M., Morgan, M., Bergsten, H., Hyldegaard, O., Skrede, S., Arnell, P., Madsen, M. B., Johansson, L., et al. (2016). "Biofilm in group A streptococcal necrotizing soft tissue infections". In: *JCI insight* 1.10.

Siemens, N. and Norrby-Teglund, A. (2017). "Shocking superantigens promote establishment of bacterial infection". In: *Proceedings of the National Academy of Sciences* 114.38, 10000–10002.

Siemens, N., Snäll, J., Svensson, M., and Norrby-Teglund, A. (2020). "Pathogenic mechanisms of streptococcal necrotizing soft tissue infections". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 127–150.

Singh, A., Thakur, N., and Sharma, A. (2016). "A review of supervised machine learning algorithms". In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee, 1310–1315.

Singh, B., Fleury, C., Jalalvand, F., and Riesbeck, K. (2012). "Human pathogens utilize host extracellular matrix proteins laminin and collagen for adhesion and invasion of the host". In: *FEMS microbiology reviews* 36.6, 1122–1180.

Škėrienė, S. and Jucevičienė, P. (2020). "Problem solving through values: A challenge for thinking and capability development". In: *Thinking Skills and Creativity* 37, 100694.

Skrede, S., Bruun, T., Rath, E., and Oppegaard, O. (2020). "Microbiological etiology of necrotizing soft tissue infections". In: *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects*, 53–71.

Sloman, S. A. and Lagnado, D. (2015). "Causality in thought". In: *Annual review of psychology* 66, 223–247.

Smeesters, P. R., McMillan, D. J., and Sriprakash, K. S. (2010). "The streptococcal M protein: a highly versatile molecule". In: *Trends in microbiology* 18.6, 275–282.

Smith, C. L., Ghosh, J., Elam, J. S., Pinkner, J. S., Hultgren, S. J., Caparon, M. G., and Ellenberger, T. (2011). "Structural basis of Streptococcus pyogenes immunity to its NAD+ glycohydrolase toxin". In: *Structure* 19.2, 192–202.

Smith, R. (2015). "A mutual information approach to calculating nonlinearity". In: *Stat* 4.1, 291–303.

Snäll, J., Linnér, A., Uhlmann, J., Siemens, N., Ibold, H., Janos, M., Linder, A., Kreikemeyer, B., Herwald, H., Johansson, L., et al. (2016). "Differential neutrophil responses to bacterial stimuli: Streptococcal strains are potent inducers of heparin-binding protein and resistin-release". In: *Scientific reports* 6.1, 1–12.

Soehnlein, O., Oehmcke, S., Rothfuchs, A., Frithiof, R., Van Rooijen, N., Mörgelin, M., Herwald, H., Lindbom, L., et al. (2008). "Neutrophil degranulation mediates severe lung damage triggered by streptococcal M1 protein". In: *European Respiratory Journal* 32.2, 405–412.

Soetaert, K. and Petzoldt, T. (2010). "Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME". In: *Journal of statistical software* 33, 1–28.

Soneson, C., Love, M. I., and Robinson, M. D. (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences". In: *F1000Research* 4.

Song, L., Langfelder, P., and Horvath, S. (2012). "Comparison of co-expression measures: mutual information, correlation, and model based indices". In: *BMC bioinformatics* 13.1, 328.
Spearman, C. (1904). "Measurement of association, Part II. Correction of 'systematic deviations'". In: *Am J Psychol* 15, 88–101.
Speidl, W. S., Kasd, S. P., Hutter, R., Katsaros, K. M., Kaun, C., Bauriedel, G., Maurer, G., Huber, K., Badimon, J. J., and Wojta, J. (2011). "The complement component C5a is present in human coronary lesions in vivo and induces the expression of MMP-1 and MMP-9 in human macrophages in vitro". In: *The FASEB Journal* 25.1, 35–44.
Sporn, M. B. (1996). "The war on cancer." In: *Lancet (London, England)* 347.9012, 1377–1381.
Springer, T. A. (1994). "Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm". In: *Cell* 76.2, 301–314.
Stanley, D., Geier, M. S., Hughes, R. J., Denman, S. E., and Moore, R. J. (2013). "Highly variable microbiota development in the chicken gastrointestinal tract". In: *PloS one* 8.12, e84290.
Stavrum, A. K., Petersen, K., Jonassen, I., and Dysvik, B. (2008). "Analysis of gene-expression data using J-Express". In: *Current Protocols in Bioinformatics* 21.1, 7–3.
Steer, A. C., Jenney, A., Kado, J., Good, M. F., Batzloff, M., Waqatakirewa, L., Mullholland, E. K., and Carapetis, J. R. (2009). "Prospective surveillance of invasive group A streptococcal disease, Fiji, 2005–2007". In: *Emerging infectious diseases* 15.2, 216.
Stekhoven, D. J. and Bühlmann, P. (2011). "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1, 112–118.
— (2012). "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1, 112–118.
Sterckx, S., Rakic, V., Cockbain, J., and Borry, P. (2016). ""You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care. data scheme and their wider relevance for biomedical research". In: *Medicine, health care and philosophy* 19, 177–190.
Stern, A. and Sorek, R. (2011). "The phage-host arms race: shaping the evolution of microbes". In: *Bioessays* 33.1, 43–51.
Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). "The mutual information: detecting and evaluating dependencies between variables". In: *Bioinformatics* 18.suppl_2, S231–S240.
Stevens, D. L. (1995). "Streptococcal toxic-shock syndrome: spectrum of disease, pathogenesis, and new concepts in treatment." In: *Emerging infectious diseases* 1.3, 69.
Stevens, D. L., Bisno, A. L., Chambers, H. F., Dellinger, E. P., Goldstein, E. J., Gorbach, S. L., Hirschmann, J. V., Kaplan, S. L., Montoya, J. G., and Wade, J. C. (2014). "Practice guidelines for the diagnosis and management of skin and soft tissue infections: 2014 update by the Infectious Diseases Society of America". In: *Clinical infectious diseases* 59.2, e10–e52.
Stevens, D. L., Bisno, A. L., Chambers, H. F., Everett, E. D., Dellinger, P., Goldstein, E. J., Gorbach, S. L., Hirschmann, J. V., Kaplan, E. L., Montoya, J. G., et al. (2005).

"Practice guidelines for the diagnosis and management of skin and soft-tissue infections". In: *Clinical Infectious Diseases* 41.10, 1373–1406.

Stevens, D. L. and Bryant, A. E. (2017). "Necrotizing soft-tissue infections". In: *New England journal of medicine* 377.23, 2253–2265.

Stevens, V. L., Wang, Y., Carter, B. D., Gaudet, M. M., and Gapstur, S. M. (2018). "Serum metabolomic profiles associated with postmenopausal hormone use". In: *Metabolomics* 14.7, 97.

Streuli, H. (1973). "Der heutige stand der kaffeechemie". In: *Association Scientifique International du Cafe, 6th International Colloquium on Coffee Chemistry, Bogota, Colombia*, 61–72.

Struck, N. S., Zimmermann, M., Krumkamp, R., Lorenz, E., Jacobs, T., Rieger, T., Wurr, S., Günther, S., Gyau Boahen, K., Marks, F., et al. (2020). "Cytokine profile distinguishes children with Plasmodium falciparum malaria from those with bacterial blood stream infections". In: *The Journal of Infectious Diseases* 221.7, 1098–1106.

Study, O. G. A. S., Kaul, R., McGeer, A., Low, D. E., Green, K., Schwartz, B., and Simor, A. E. (1997). "Population-based surveillance for group A streptococcal necrotizing fasciitis: clinical features, prognostic indicators, and microbiologic analysis of seventy-seven cases". In: *The American journal of medicine* 103.1, 18–24.

Suarez-Diez, M. and Saccenti, E. (2015). "Effects of sample size and dimensionality on the performance of four algorithms for inference of association networks in metabonomics". In: *Journal of proteome research* 14.12, 5119–5130.

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., et al. (2015). "Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools". In: *Nucleic acids research* 44.D1, D463–D470.

Sudarsky, L. A., Laschinger, J. C., Coppa, G. F., and Spencer, F. C. (1987). "Improved results from a standardized approach in treating patients with necrotizing fasciitis." In: *Annals of surgery* 206.5, 661.

Suijker, J., Vries, A. de, Jong, V. M. de, Schepers, T., Ponsen, K. J., and Halm, J. A. (2020). "Health-related quality of life is decreased after necrotizing soft-tissue infections". In: *journal of surgical research* 245, 516–522.

Sullivan, L. M., Weinberg, J., and Keaney Jr, J. F. (2016). "Common statistical pitfalls in basic science research". In: *Journal of the American Heart Association* 5.10, e004142.

Suttle, C. A. (2007). "Marine viruses—major players in the global ecosystem". In: *Nature reviews microbiology* 5.10, 801–812.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). "UniRef: comprehensive and non-redundant UniProt reference clusters". In: *Bioinformatics* 23.10, 1282–1288.

Swartz, T. H., Palermo, A.-G. S., Masur, S. K., and Aberg, J. A. (2019). "The science and value of diversity: closing the gaps in our understanding of inclusion and diversity". In: *The Journal of infectious diseases* 220.Supplement_2, S33–S41.

Tan, L. K., Reglinski, M., Teo, D., Reza, N., Lamb, L. E., Nageshwaran, V., Turner, C. E., Wikstrom, M., Frick, I.-M., Bjorck, L., et al. (2021). "Vaccine-induced, but

not natural immunity, against the Streptococcal inhibitor of complement protects against invasive disease". In: *npj Vaccines* 6.1, 62.
Tang, B. M., Huang, S. J., and McLean, A. S. (2010). "Genome-wide transcription profiling of human sepsis: a systematic review". In: *Critical care* 14, 1–11.
Tavassoly, I., Goldfarb, J., and Iyengar, R. (2018). "Systems biology primer: the basic methods and approaches". In: *Essays In Biochemistry*. DOI: 10.1042/EBC20180003.
Team, R. C. et al. (2016). "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria". In: *http://www. R-project. org/*.
Tedde, V., Rosini, R., and Galeotti, C. L. (2016). "Zn2+ Uptake in Streptococcus pyogenes: Characterization of adcA and lmb Null Mutants". In: *PLoS One* 11.3, e0152835.
Terao, Y., Kawabata, S., Kunitomo, E., Murakami, J., Nakagawa, I., and Hamada, S. (2008). "Fba, a novel fibronectin-binding protein from Streptococcus pyogenes, promotes bacterial entry into epithelial cells, and the fba gene is positively transcribed under the Mga regulator". In: *Molecular microbiology* 42.1, 75–86.
Terao, Y., Kawabata, S., Kunitomo, E., Nakagawa, I., and Hamada, S. (2002). "Novel laminin-binding protein of Streptococcus pyogenes, Lbp, is involved in adhesion to epithelial cells". In: *Infection and immunity* 70.2, 993–997.
Terrier, L.-M., Lévêque, M., and Amelot, A. (2019). "Brain lobotomy: a historical and moral dilemma with no alternative?" In: *World neurosurgery* 132, 211–218.
Thänert, R., Itzek, A., Hoßmann, J., Hamisch, D., Madsen, M. B., Hyldegaard, O., Skrede, S., Bruun, T., Norrby-Teglund, A., Oppegaard Oddvar 4 Rath Eivind 4 Nedrebø Torbjørn 4 Arnell Per 8 Rosen Anders 8 Polzik Peter 3 Hansen Marco Bo 3 Svensson Mattias 6 Snäll Johanna 6 Karlsson Ylva 9 Nekludov Michael 10, I. study group, et al. (2019). "Molecular profiling of tissue biopsies reveals unique signatures associated with streptococcal necrotizing soft tissue infections". In: *Nature communications* 10.1, 3846.
Thévenot, E. A., Roux, A., Xu, Y., Ezan, E., and Junot, C. (2015). "Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses". In: *Journal of proteome research* 14.8, 3322–3335.
Thomas, D. J. and Bralower, T. J. (2005). "Sedimentary trace element constraints on the role of North Atlantic Igneous Province volcanism in late Paleocene–early Eocene environmental change". In: *Marine Geology* 217.3-4, 233–254.
Thulin, P., Johansson, L., Low, D. E., Gan, B. S., Kotb, M., McGeer, A., and Norrby-Teglund, A. (2006). "Viable group A streptococci in macrophages during acute soft tissue infection". In: *PLoS medicine* 3.3, e53.
Tilley, L. (2015). "Accommodating difference in the prehistoric past: Revisiting the case of Romito 2 from a bioarchaeology of care perspective". In: *International Journal of Paleopathology* 8, 64–74.
Tokunaga, R., Zhang, W., Naseem, M., Puccini, A., Berger, M. D., Soni, S., McSkane, M., Baba, H., and Lenz, H.-J. (2018). "CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation–a target for novel cancer therapy". In: *Cancer treatment reviews* 63, 40–47.
Tong, W. H., Pavey, C., O'Handley, R., and Vyas, A. (2021). "Behavioral biology of Toxoplasma gondii infection". In: *Parasites & Vectors* 14, 1–6.

Trinkaus, E. and Villotte, S. (2017). "External auditory exostoses and hearing loss in the Shanidar 1 Neandertal". In: *PLoS One* 12.10, e0186684.
Tripathi, S., Dehmer, M., and Emmert-Streib, F. (2014). "NetBioV: an R package for visualizing large network data in biology and medicine". In: *Bioinformatics* 30.19, 2834–2836.
Trudeau, R. J. (2013). *Introduction to graph theory*. Courier Corporation.
Truong, D., Copeland, J. W., and Brumell, J. H. (2014). "Bacterial subversion of host cytoskeletal machinery: hijacking formins and the Arp2/3 complex". In: *BioEssays* 36.7, 687–696.
Tsai, S., Hardison, N. E., James, A. H., Motsinger-Reif, A. A., Bischoff, S. R., Thames, B. H., and Piedrahita, J. A. (2011). "Transcriptional profiling of human placentas from pregnancies complicated by preeclampsia reveals disregulation of sialic acid acetylesterase and immune signalling pathways". In: *Placenta* 32.2, 175–182.
Tso, D. K. and Singh, A. K. (2018). "Necrotizing fasciitis of the lower extremity: imaging pearls and pitfalls". In: *The British Journal of Radiology* 91.1088, 20180093.
Tyagi, V., Hanoch, Y., Hall, S. D., Runco, M., and Denham, S. L. (2017). "The risky side of creativity: Domain specific risk taking in creative individuals". In: *Frontiers in psychology* 8, 145.
Urbina, T., Canoui-Poitrine, F., Hua, C., Layese, R., Alves, A., Ouedraogo, R., Bosc, R., Sbidian, E., Chosidow, O., Dessap, A. M., et al. (2021). "Long-term quality of life in necrotizing soft-tissue infection survivors: a monocentric prospective cohort study". In: *Annals of Intensive Care* 11.1, 1–11.
Valderrama, J. A. and Nizet, V. (2018). "Group A Streptococcus encounters with host macrophages". In: *Future Microbiology* 13.1, 119–134.
Van der Maaten, L. and Hinton, G. (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.
Van Engelen, T. S., Wiersinga, W. J., Scicluna, B. P., and Poll, T. van der (2018). "Biomarkers in sepsis". In: *Critical care clinics* 34.1, 139–152.
Vapnik, V., Golowich, S., and Smola, A. (1996). "Support vector method for function approximation, regression estimation and signal processing". In: *Advances in neural information processing systems* 9.
Vasey, P. L. (2002). "Sexual partner preference in female Japanese macaques". In: *Archives of Sexual Behavior* 31, 51–62.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.
Veen, S. van der and Abee, T. (2010). "HrcA and DnaK are important for static and continuous-flow biofilm formation and disinfectant resistance in Listeria monocytogenes". In: *Microbiology* 156.12, 3782–3790.
Vega, L. A., Malke, H., and McIver, K. S. (2022). "Virulence-related transcriptional regulators of Streptococcus pyogenes". In.
Vehlow, C., Beck, F., and Weiskopf, D. (2015). "The State of the Art in Visualizing Group Structures in Graphs." In: *EuroVis (STARs)*, 21–40.
— (2017). "Visualizing group structures in graphs: A survey". In: *Computer Graphics Forum*. Vol. 36. 6. Wiley Online Library, 201–225.
Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). "Interactome networks and human disease". In: *Cell* 144.6, 986–998.

Vidulin, V., Bohanec, M., and Gams, M. (2014). "Combining human analysis and machine data mining to obtain credible data relations". In: *Information Sciences* 288, 254–278.

Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L., Turano, P., and Luchinat, C. (2019). "High-throughput metabolomics by 1D NMR". In: *Angewandte Chemie International Edition* 58.4, 968–994.

Vignoli, A., Tenori, L., Giusti, B., Valente, S., Carrabba, N., Balzi, D., Barchielli, A., Marchionni, N., Gensini, G. F., Marcucci, R., et al. (2020). "Differential network analysis reveals metabolic determinants associated with mortality in acute myocardial infarction patients and suggest potential mechanisms underlying different clinical scores used to predict death". In: *Journal of Proteome Research*.

Vignoli, A., Tenori, L., Luchinat, C., and Saccenti, E. (2017). "Age and sex effects on plasma metabolite association networks in healthy subjects". In: *Journal of proteome research* 17.1, 97–107.

Vincent, J. .-., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. G. (1996). *The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix)*.

Viti, A., Terzi, A., and Bertolaccini, L. (2015). "A practical overview on probability distributions". In: *Journal of Thoracic Disease* 7.3, E7.

Vlaminckx, B., Van Pelt, W., Schouls, L., Van Silfhout, A., Elzenaar, C., Mascini, E., Verhoef, J., and Schellekens, J. (2004). "Epidemiological features of invasive and noninvasive group A streptococcal disease in the Netherlands, 1992–1996". In: *European Journal of Clinical Microbiology and Infectious Diseases* 23, 434–444.

Vogl, T., Eisenblätter, M., Völler, T., Zenker, S., Hermann, S., Van Lent, P., Faust, A., Geyer, C., Petersen, B., Roebrock, K., et al. (2014). "Alarmin S100A8/S100A9 as a biomarker for molecular imaging of local inflammatory activity". In: *Nature communications* 5.1, 4593.

Vogl, T., Tenbrock, K., Ludwig, S., Leukert, N., Ehrhardt, C., Van Zoelen, M. A., Nacken, W., Foell, D., Van der Poll, T., Sorg, C., et al. (2007). "Mrp8 and Mrp14 are endogenous activators of Toll-like receptor 4, promoting lethal, endotoxin-induced shock". In: *Nature medicine* 13.9, 1042–1049.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S., Myles, P., et al. (2018). "Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness". In: *arXiv preprint arXiv:1812.10404*.

Wang, R.-S., Maron, B. A., and Loscalzo, J. (2015). "Systems medicine: evolution of systems biology from bench to bedside". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 7.4, 141–161.

Wang, S.-M., Lu, I.-H., Lin, Y.-L., Lin, Y.-S., Wu, J.-J., Chuang, W.-J., Lin, M. T., and Liu, C.-C. (2008). "The severity of Streptococcus pyogenes infections in children is significantly associated with plasma levels of inflammatory cytokines". In: *Diagnostic microbiology and infectious disease* 61.2, 165–169.

Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., and Ma, J. (2018). "S100A8/A9 in Inflammation". In: *Frontiers in immunology* 9, 1298.

Wang, Y., Shen, Q., Archambault, D., Zhou, Z., Zhu, M., Yang, S., and Qu, H. (2015). "Ambiguityvis: Visualization of ambiguity in graph layouts". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, 359–368.

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., et al. (2017). "Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center". In: *Nucleic acids research* 45.D1, D535–D542.

Weaver, T. E. (2008). "Enhancing multiple disciplinary teamwork". In: *Nursing outlook* 56.3, 108–114.

Wedel, M. J. (2011). "A monument of inefficiency: The presumed course of the recurrent laryngeal nerve in sauropod dinosaurs". In: *Acta Palaeontologica Polonica* 57.2, 251–256.

Wehrens, R. and Franceschi, P. (2012). "Meta-Statistics for Variable Selection: The R Package BioMark". In: *Journal of Statistical Software, Articles* 51.10, 1–18. DOI: 10.18637/jss.v051.i10. URL: https://www.jstatsoft.org/v051/i10.

Wei, R., Wang, J., Jia, E., Chen, T., Ni, Y., and Jia, W. (2018). "GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies". In: *PLoS computational biology* 14.1, e1005973.

Weiler, H. and Isermann, B. (2003). "Thrombomodulin". In: *Journal of thrombosis and haemostasis* 1.7, 1515–1524.

Welch, M. D. and Way, M. (2013). "Arp2/3-mediated actin-based motility: a tail of pathogen abuse". In: *Cell host & microbe* 14.3, 242–255.

Weller-Stuart, T., De Maayer, P., and Coutinho, T. (2017). "Pantoea ananatis: genomic insights into a versatile pathogen". In: *Molecular plant pathology* 18.9, 1191–1198.

Westman, J., Chakrakodi, B., Snäll, J., Mörgelin, M., Bruun Madsen, M., Hyldegaard, O., Neumann, A., Frick, I.-M., Norrby-Teglund, A., Björck, L., et al. (2018). "Protein SIC secreted from Streptococcus pyogenes forms complexes with extracellular histones that boost cytokine production". In: *Frontiers in Immunology* 9, 236.

White, B., Schmidt, M., Murphy, C., Livingstone, W., O'toole, D., Lawler, M., O'Neill, L., Kelleher, D., Schwarz, H., and Smith, O. (2000). "Activated protein C inhibits lipopolysaccharide-induced nuclear translocation of nuclear factor κB (NF-κB) and tumour necrosis factor α (TNF-α) production in the THP-1 monocytic cell line". In: *British journal of haematology* 110.1, 130–134.

White, H. A. (2020). "Thinking "outside the box": Unconstrained creative generation in adults with attention deficit hyperactivity disorder". In: *The Journal of Creative Behavior* 54.2, 472–483.

White, P. A. (2014). "Singular clues to causality and their use in human causal judgment". In: *Cognitive science* 38.1, 38–75.

Whitfield, K. and Reid, C. (2004). "Assumptions, ambiguities, and possibilities in interdisciplinary population health research". In: *Canadian Journal of Public Health* 95, 434–436.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1, 1–9.

Wille, A. and Bühlmann, P. (2006). "Low-order conditional independence graphs for inferring genetic networks". In: *Statistical applications in genetics and molecular biology* 5.1.
Wilson, B. (1952). "Necrotizing faciitis". In: *Am J Surg* 18, 416–431.
Wilson, D. N. (2014). "Ribosome-targeting antibiotics and mechanisms of bacterial resistance". In: *Nature Reviews Microbiology* 12.1, 35–48.
Wolkenhauer, O., Auffray, C., Jaster, R., Steinhoff, G., and Dammann, O. (2013). "The road from systems biology to systems medicine". In: *Pediatric research* 73.2, 502–507.
Wong, C.-H., Chang, H.-C., Pasupathy, S., Khin, L.-W., Tan, J.-L., and Low, C.-O. (2003). "Necrotizing fasciitis: clinical presentation, microbiology, and determinants of mortality". In: *JBJS* 85.8, 1454–1460.
Wong, C.-H., Khin, L.-W., Heng, K.-S., Tan, K.-C., and Low, C.-O. (2004). "The LRINEC (Laboratory Risk Indicator for Necrotizing Fasciitis) score: a tool for distinguishing necrotizing fasciitis from other soft tissue infections". In: *Critical care medicine* 32.7, 1535–1541.
Wood, R. E., Beckmann, N., and Rossiter, J. R. (2011). "Management humor: Asset or liability?" In: *Organizational Psychology Review* 1.4, 316–338.
Woodbury, R. and Haldenwang, W. (2003). "HrcA is a negative regulator of the dnaK and groESL operons of Streptococcus pyogenes". In: *Biochemical and biophysical research communications* 302.4, 722–727.
Wright, S. W., Kaewarpai, T., Lovelace-Macon, L., Ducken, D., Hantrakun, V., Rudd, K. E., Teparrukkul, P., Phunpang, R., Ekchariyawat, P., Dulsuk, A., et al. (2021). "A 2-biomarker model augments clinical prediction of mortality in melioidosis". In: *Clinical Infectious Diseases* 72.5, 821–828.
Wu, L., Neskovic, P., Reyes, E., Festa, E., and William, H. (2007). "Classifying n-back EEG data using entropy and mutual information features." In: *ESANN*, 61–66.
Xia, P., Lian, S., Wu, Y., Yan, L., Quan, G., and Zhu, G. (2021). "Zinc is an important inter-kingdom signal between the host and microbe". In: *Veterinary Research* 52, 1–14.
Yabluchanskiy, A., Ma, Y., Iyer, R. P., Hall, M. E., and Lindsey, M. L. (2013). "Matrix metalloproteinase-9: Many shades of function in cardiovascular disease". In: *Physiology* 28.6, 391–403.
Yamakawa, K., Aihara, M., Ogura, H., Yuhara, H., Hamasaki, T., and Shimazu, T. (2015). "Recombinant human soluble thrombomodulin in severe sepsis: a systematic review and meta-analysis". In: *Journal of Thrombosis and Haemostasis* 13.4, 508–519.
Yamakawa, K., Ogura, H., Fujimi, S., Morikawa, M., Ogawa, Y., Mohri, T., Nakamori, Y., Inoue, Y., Kuwagata, Y., Tanaka, H., et al. (2013). "Recombinant human soluble thrombomodulin in sepsis-induced disseminated intravascular coagulation: a multicenter propensity score analysis". In: *Intensive Care Medicine* 39, 644–652.
Yan, F., Robert, M., and Li, Y. (2017). "Statistical methods and common problems in medical or biomedical science research". In: *International journal of physiology, pathophysiology and pharmacology* 9.5, 157.
Yang, L., Manithody, C., Walston, T. D., Cooper, S. T., and Rezaie, A. R. (2003). "Thrombomodulin enhances the reactivity of thrombin with protein C inhibitor by pro-

viding both a binding site for the serpin and allosterically modulating the activity of thrombin". In: *Journal of Biological Chemistry* 278.39, 37465–37470.

Yang, L., Froio, R. M., Sciuto, T. E., Dvorak, A. M., Alon, R., and Luscinskas, F. W. (2005). "ICAM-1 regulates neutrophil adhesion and transcellular migration of TNF-α-activated vascular endothelium under flow". In: *Blood* 106.2, 584–592.

Ye, W., Chen, G., Li, X., Lan, X., Ji, C., Hou, M., Zhang, D., Zeng, G., Wang, Y., Xu, C., et al. (2020). "Dynamic changes of D-dimer and neutrophil-lymphocyte count ratio as prognostic biomarkers in COVID-19". In: *Respiratory research* 21.1, 1–7.

Yen, Z.-S., Wang, H.-P., Ma, H.-M., Chen, S.-C., and Chen, W.-J. (2002). "Ultrasonographic screening of clinically-suspected necrotizing fasciitis". In: *Academic emergency medicine* 9.12, 1448–1451.

You, Y., Liang, D., Wei, R., Li, M., Li, Y., Wang, J., Wang, X., Zheng, X., Jia, W., and Chen, T. (2019). "Evaluation of metabolite-microbe correlation detection methods". In: *Analytical biochemistry* 567, 106–111.

Yu, C. and Yao, W. (2017). "Robust linear regression: A review and comparison". In: *Communications in Statistics-Simulation and Computation* 46.8, 6261–6282.

Yura, T. and Nakahigashi, K. (1999). "Regulation of the heat-shock response". In: *Current opinion in microbiology* 2.2, 153–158.

Zeerleder, S., Hack, C. E., and Wuillemin, W. A. (2005). "Disseminated intravascular coagulation in sepsis". In: *Chest* 128.4, 2864–2875.

Zelezniak, A., Sheridan, S., and Patil, K. R. (2014). "Contribution of Network Connectivity in Determining the Relationship between Gene Expression and Metabolite Concentration Changes". In: *PLOS Computational Biology* 10.4, 1–12. DOI: 10.1371/journal.pcbi.1003572. URL: https://doi.org/10.1371/journal.pcbi.1003572.

Zhang, B., Tian, Y., and Zhang, Z. (2014). "Network biology in medicine and beyond". In: *Circulation: Cardiovascular Genetics* 7.4, 536–547.

Zhang, W., Zeng, T., Liu, X., and Chen, L. (2015). "Diagnosing phenotypes of single-sample individuals by edge biomarkers". In: *Journal of molecular cell biology* 7.3, 231–241.

Zhang, W., Jang, S., Jonsson, C. B., and Allen, L. J. (2019). "Models of cytokine dynamics in the inflammatory response of viral zoonotic infectious diseases". In: *Mathematical medicine and biology: a journal of the IMA* 36.3, 269–295.

Zhao, J., Zhou, Y., Zhang, X., and Chen, L. (2016). "Part mutual information for quantifying direct associations in networks". In: *Proceedings of the National Academy of Sciences* 113.18, 5130–5135.

Zheng, L., Chen, Z., Itzek, A., Herzberg, M. C., and Kreth, J. (2012). "CcpA regulates biofilm formation and competence in Streptococcus gordonii". In: *Molecular oral microbiology* 27.2, 83–94.

Zheng, X., Huang, F., Zhao, A., Lei, S., Zhang, Y., Xie, G., Chen, T., Qu, C., Rajani, C., Dong, B., et al. (2017). "Bile acid is a significant host factor shaping the gut microbiome of diet-induced obese mice". In: *BMC biology* 15.1, 120.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). "Graph neural networks: A review of methods and applications". In: *AI open* 1, 57–81.

Zhou, Y., Stoudenmire, E. M., and Waintal, X. (2020). "What limits the simulation of quantum computers?" In: *Physical Review X* 10.4, 041038.

Zimmerman, D. W., Zumbo, B. D., and Williams, R. H. (2003). "Bias in estimation and hypothesis testing of correlation". In: *Psicológica* 24.1.
Zonneveld, R., Martinelli, R., Shapiro, N. I., Kuijpers, T. W., Plötz, F. B., and Carman, C. V. (2014). "Soluble adhesion molecules as markers for sepsis and the potential pathophysiological discrepancy in neonates, children and adults". In: *Critical care* 18, 1–14.

Author Affiliations

1. Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands.
2. Department of Medicine, Division for infectious diseases, Haukeland University Hospital, Bergen, Norway
3. Department of Clinical Science, University of Bergen, Bergen, Norway
4. Center for Infectious Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Huddinge, Sweden
5. Department of Anesthesia, Centre of Head and Orthopaedics, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
6. Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
7. Bioprocess Engineering, Wageningen University & Research, Wageningen, The Netherlands
8. Lifeglimmer GmbH, Markelstraße 38, 12163, Berlin, Germany
9. Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands
10. Department of Medicine Huddinge, Karolinska Institute, Stockholm, Sweden.
11. Department of Infectious Diseases, Karolinska University Hospital, Stockholm, Sweden.
12. Functional Area of Emergency Medicine, Karolinska University Hospital, Stockholm, Sweden.
13. Department of Anaesthesia and Intensive Care, Sahlgrenska University Hospital, Gothenburg, Sweden
14. Department of Anaesthesia, Surgical Services and Intensive Care, Karolinska Institute, Karolinska University Hospital, Stockholm, Sweden.
15. Childhood Cancer Research Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
16. Department of Infectious Diseases, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway
17. Centre of Molecular Inflammation Research, Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

18. Genomic Clinical Services and Data Center, Kingmed Diagnostic, Shanghai, China

19. Department of Basic Medical Sciences, University of Lleida, Lleida, Spain

20. Department of Environmental Science, Wageningen University and Research, Wageningen, The Netherlands.

21. Department of Social Sciences, Wageningen University & Research, The Netherlands

22. Bioinformatics Group, Wageningen University & Research, The Netherlands

List of Publications

Manuscripts included in this thesis

- **Jahagirdar, S.**, & Saccenti, E. (2020). On the Use of Correlation and MI as a Measure of Metabolite—Metabolite Association for Network Differential Connectivity Analysis. Metabolites, 10(4), 171.
- **Jahagirdar, S.**, & Saccenti, E. (2021). Evaluation of Single Sample Network Inference Methods for Metabolomics-Based Systems Medicine. Journal of Proteome Research, 20(1), 932–949.
- Palma Medina, L. M., Rath, E.*, **Jahagirdar, S.***, Bruun, T., Madsen, M. B., Strålin, K., Unge, C., Hansen, M. B., Arnell, P., Nekludov, M., Hyldegaard, O., Lourda, M., Santos, V. A. P. M. dos, Saccenti, E., Skrede, S., Svensson, M., & Norrby-Teglund, A. (2021). Discriminatory plasma biomarkers predict specific clinical phenotypes of necrotizing soft-tissue infections. Journal of Clinical Investigation, 131(14). ***Contributed equally**
- **Jahagirdar, S.**, Morris, L., Benis, N., Oppegaard, O., Svenson, M., Hyldegaard, O., Skrede, S., Norrby-Teglund, A., Bruun, T., Rath, E., Nedrebø, T., Arnell, P., Rosen, A., Hedetoft, M., Madsen, M. B., Svensson, M., Snäll, J., Karlsson, Y., Nekludov, M., Santos, V. A. P. M. dos, & Saccenti, E. (2022). Analysis of host-pathogen gene association networks reveals patient-specific response to streptococcal and polymicrobial necrotising soft tissue infections. BMC Medicine, 20(1), 173.
- Rath, E., Palma Medina, L. M.*, **Jahagirdar, S.***, Mosevoll, K. A., Damås, J. K., Madsen, M. B., Svensson, M., Hyldegaard, O., Martins dos Santos, V. A. P., Saccenti, E., Norrby-Teglund, A., Skrede, S., & Bruun, T. (2023). Systemic immune activation profiles in streptococcal necrotizing soft tissue infections: A prospective multicenter study. Clinical Immunology, 249, 109276. ***Contributed equally**
- **Jahagirdar, S.**, Balder, Y., Mosevoll, K. A., Oppegaard, O., Reikvam, H., Bruun, T., Svenson, M., Hyldegaard, O., Norrby-Telgund, A, Santos, V. A. P. M. dos, Skrede, S., & Saccenti, E. Multiomics analyses of mixed data from NSTI patients reveal dependencies with clinical variables. In preparation for submission
- **Jahagirdar, S.**, Lamers, M.*, Kuang, C.*, Basallo Clariana, O.*, Palma Medina, L. M., Mosevoll, K. A., Rath, E., Bruun, T., Svensson, M., Hyldegaard, O., Norrby-Teglund, A., Skrede, S., Santos, V. A. P. M. dos, & Saccenti, E., Analysis of plasma analyte ratio and model parameters reveal pair-wise relationships of cytokines in NSTI. In preparation for submission. *Contributed equally
- Koetsier, R., Setty, S., Saccent, E.[$], Van de Leemput, I.[$], Suarez-Diez, M.[$], Santos, V. A. P. M. dos.[$], & **Jahagirdar, S.** Automated network visualisation software: anvis. In preparation for submission. $Author names organised alphabetically.

- **Jahagirdar, S.***, Setty, S.*, Robaey, Z., & Santos, V. A. P. M. dos. Problem-solving in multi-/inter-/trans-disciplinary environments from an early career researcher perspective. In preparation for submission

Manuscripts not included in the thesis

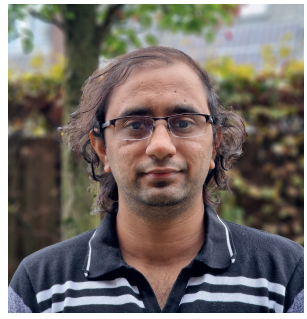
- **Jahagirdar, S.**, Suarez-Diez, M., & Saccenti, E. (2019). Simulation and Reconstruction of Metabolite–Metabolite Association Networks Using a Metabolic Dynamic Model and Correlation Based Algorithms. Journal of Proteome Research, 18(3), 1099–1113.
- Wang, X., **Jahagirdar, S.** Kemp, B., Saccenti, E., & Van Knegsel, A. Ability of plasma and milk indicators to classify diet, dry period length, and lactation week of dairy cows using a machine learning approach. In preparation for submission
- Wang, X., **Jahagirdar, S.**, Bakker, W., Lute, C., Kemp, B., Van Knegsel, A., & Saccenti, E. Classification of cows to a lipogenic or glucogenic diet based on metabolite ratios using a decision-tree based algorithm. In preparation for submission
- Setty, S. Boot, A., **Jahagirdar, S.**, Sluijs, A., & Dijksta, H. Understanding causal skills using a global carbon model. In preparation for submission.

Overview of Completed Training Activities

	Organising Institute	Year
Discipline specific activities		
PerMIT Kick-off	LifeGlimmer	2019
Data Visualisation Master Class	DCS	2019
BioSB conference + PhD retreat	BioSB	2020
PerAID 1st annual	Norwegian University of Science & Technology	2020
PerMIT 1st annual	WUR	2020
Modelling & Optimisation for Bioinformatics & Systems Biology	BioSB	2021
Algorithms for biological networks	BioSB	2021
BioSB conference + PhD retreat	BioSB	2021
PerAID/PerMIT review	University of Bergen	2021
PerAID 2nd annual	Karolinska Institute	2021
PerMIT 2nd annual	LifeGlimmer/WUR	2021
PerAID/PerMIT review	Norwegian University of Science & Technology	2022
PerAID 3rd annual	Karolinska Institute	2022
PerMIT 3rd annual	Karolinska Institute	2022
PerMIT-PerAID consortium scientific meet	Righospitalet	2023
General Courses		
VLAG PhD week	VLAG	2019
Project and Time Managemnent	WGS	2021
Phd competence assessment	WGS	2021
Intensive writing week	WGS	2022
Mobilising your scientific network	WGS	2022
Critical Thinking and Argumentation	WGS	2022
Essentials of Scientific Writing and Presentation	WGS	2022
Career Orientation	WGS	2022
Assisting in teaching and supervision activities		
WUR MSc courses: SSB-30306 (Molecular Systems Biology), SSB-31312 (Toolbox)		2021-2023
Supervising Msc Students		2020-2022

Other activities		
Preparation of research proposal	SSB	2020
SSB Workgroup meetings	SSB	2019-2023
DS/AI Fellowship (Received)	WDCC	2022

About the Author



Sanjeevan's tenacious professional journey is a testament to his unwavering commitment to the principles of logic and reason, his insatiable thirst for knowledge, and his resolute dedication to moral and ethical values. Sanjeevan was born on the 13th of April 1992 in Pune, where he received his early education. His passion for Science was clearly evident as he enthusiastically wrote projects spanning over 100 handwritten pages on diverse topics for no credit gain. Upon completing his secondary education, he pursued a Bachelor's degree in Biotechnology at the University of Pune (now known as Savitribai Phule Pune University). It was during this period that he began to recognize the vast potential of computational and systems biology. Despite the absence of formal coursework on systems biology within his program, he avidly delved into the subject by independently studying systems biology with free material available on the internet such as Uri Alon's lectures. Realising the need for programming knowledge in this field, he soon began auditing online courses like CS50x. This culminated in him doing a bachelor's thesis in bioinformatics with the National Chemical Laboratory (NCL). Following this, he moved to Wageningen to do his Master's degree in (Medical) Biotechnology. Upon articulating his academic interests, his study advisor directed him toward the course "toolbox in systems and synthetic biology". Here, he was granted full autonomy to undertake a computational project aligned with his research interests. Here he worked on network inference methods and this was the start of his long affiliation with the Laboratory of Systems and Synthetic Biology (SSB). He continued the work on network inference in SSB for his master's thesis for which he was nominated for the best thesis award. He followed this with a second thesis at SSB where he modelled the biochemical reactions of a bacteria. After graduating with his master's degree in 2019, he chose to accept the offer to do a PhD in systems medicine with SSB at Wageningen University & Research. During his PhD, he used computational approaches to understand and explore the underlying biological mechanisms in Necrotising Soft Tissue Infections. During this period he maintained successful cross-functional collaborations with biologists, micro-biologists, medical doctors, and bioinformaticians. He further received an internal grant/fellowship to create a network visualisation software. The work done in his PhD has resulted in this thesis. In August 2023, he did a very short post-doc in SSB to help design teaching material for the Modelling in Systems Biology course. He now turns his attention to seek a new challenge that could quench his ever growng thirst for knowledge.

Acknowledgements

Now that you have read through the entire book (or not) and have turned over to this page, you know why you are here! It's time for the monologue! I have to acknowledge that pursuing this PhD has proven to be one of the easier challenges in my life, in stark contrast to the earlier stages on my educational ladder. I thoroughly enjoyed the entire process from the start to finish and this experience owes all of its richness to the interactions I've had with all of you. So, without further ado, I would like to seize this moment (space) to express my sincere gratitude.

Without a doubt, I wouldn't be here composing these words without your presence in my life: ~~Princess Carolyn~~ **Shruti Setty**. You are a strong and independent woman who decided to walk into my life and improve it forever. Without you and your help, I wouldn't have a bachelor's degree, a master's degree and definitely not a PhD. It is all well and good pretending we are objective robots doing Science, but let's be honest so many events took place in our personal lives during these 4 years and it is impossible to say none of them had any influence on our work. At the start of this PhD, you had fought your way out of a dilapidating environment that was attempting to control your behavior, information, thoughts, emotions, career, and relationships only to be confronted head on with the most insane proposition wherein you could either have a relationship with me or with your parents. I am very grateful you chose to build your happy life with me. The PhDs were our jobs, the jobs were our independence, we didn't have the privilege of taking 6 months or a year off to process stuff, it had to be done after work hours. Emotional drama, games, harassment, and legal threats later, here we both are on track to produce some excellent PhD theses. I am proud of you, I am proud of us, and I am proud of the environment we built for us. I am proud of the fact that you have constantly created the opportunities that have led to your success. As I have often said, people may have fast cars, but you have the ability to build a road where there is none and then drive on it. If I start jotting down every way that you have helped me deal with my emotions, depressions, empowered me and made me accept that I am infact very hardworking and not lazy despite the words and impact of "the one who shall not be named", then this would soon become a very long essay. I am thankful for everything you have done for me and I love you very much. Needless to say, I have dedicated this book to you and the words on the first page are in reference to you and your life.

Edo, I first met you during the toolbox course, where you were my supervisor. You offered me complete freedom to design any computational project, following which you were excited to publish the results. This led to my master's thesis and eventually to this PhD. Our collaboration has yielded several published manuscripts and even more in the pipeline ready to be submitted. This is no mean feat. I am thankful to you for the recognition of my potential, active promotion of my work, and the constant encouragement and celebration of my artistic expressions. I have grown a lot as a researcher in these years, and I think you have also grown a lot as a supervisor. Thank you for the supervision, but also for sharing this long journey with me.

Vitor, as my promoter, you have been one of the first person in the position of power that has promised to support any scientific avenue that I wanted to explore. This means a lot to me, given my history of finding myself in positions where I was not offered this freedom, especially when I tended to thrive on it. You pushed for me to be more independent and collaborative in my efforts and gave me space to operate. "Planned thinking time" is what you called it. You helped me improve my communication style and also in general my acumen for all things outside of technical/scientific discussions. I also learned a thing or two about networking and arbitrating in difficult situations from watching you, not sure I can replicate them though. But even more importantly, you were always a good listening ear to both my ideas and difficulties.

Maria, you have been a fantastic influence in my journey. Your office door has always been open for me to discuss any technical, personal or professional problems I may be facing. You have always been completely honest in your communication, always choosing to paint the complete nuanced picture. It was in one such discussion, that you encouraged me to apply for the DS/AI fellowship resulting in chapter 10 of this thesis. It has been one of the highlight experiences in my PhD. You also seem to manage the perfect balance between human empathy and critical comments on the work. In one our meetings, you told me, that I was the only one who needed to walk away happy from the meeting. I will carry this advice forever, especially when I find myself in supervisory/leadership roles. Thank you for walking to my office for the sole purpose of asking how I was doing from time to time. It meant a lot. I wish you the best for your future and for your vision of SSB.

Sara M, I know how busy you are with the end of your own PhD looming in site, so I'm very happy and grateful that you agreed to be my paranymph. I am happy to have shared an office with you for half of my time in SSB. We also shared an office for a very brief time when I was finishing my master's thesis and you were starting yours, however we only got to know each other mostly after COVID. I was happy to see you in the office regularly when most of SSB was a ghost town not completely recovered from "Work-from-home". You are an extremely empathetic and empowering person who always saw and assumed the good in others. It's funny that we are such different personalities that found some middle ground through our PhD experiences where I learnt to be a little more organised and you learnt to embrace the uncertainty a little. Thank you for being a very empathetic and empowering ear when I experienced any difficulties in my PhD. I really appreciated it and often left home feeling much better about myself.

Marco, we shared an office for a short time, but already you were the hardest working person there. You always came early morning and left late evening. I am happy to see you take long holidays nowadays even though you still work very hard. It was always good talking to you, we often connected on many different subjects and more often than not talked way too much. We often found parallels in some aspects of life in Mexico and India. Perhaps it was a good thing that we ended up in different offices towards the end of my PhD, more working and less talking till late in the evening. I am also thankful to you for being a patient ear if I was experiencing a stressful situation. Thank you for being my paranymph.

Efsun, you are a very kind and eagerly helpful person. I only got to know you after COVID as I attempted to steal your office desk (not really, I was just ignorant about

it). You often went above and beyond in being helpful and when I asked your advise, you guided me through the whole process of sponsoring my parents' visa, which went flawlessly. Thank you for being my paranymph and thank you for introducing me to Manti, Turkish coffee, Raki and other Turkish delicacies.

When I started my PhD, the first person I got an email from was you, **Laura**. Soon, we had our first meeting and you explained to me every single thing about the project, starting from NSTI, what measurements, how the measurements were taken, how the data is organised, what the 0s, 1s and 2s and everything else meant. A few days later you were in Wageningen with some Swedish Chokladbollar. I was told that I was going to be collaborating with a biologist, but you turned out to be a secret mathematician at heart as well. Thank you for the great collaboration and teaching me so many things and being such a key factor in so many of the projects in this thesis.

Eivind, it's a shame that we never actually met in real life even though we worked together constantly throughout the 4 years. I blame COVID-19 for this. It has been fantastic working with you. Congratulations on your own PhD. I just remember how busy you used to get, and joining online meetings for literally 15 minute between tending to patients. Thank you for patiently explaining so many medical concepts to me and an excellent collaboration.

Lorna, thank you for working with me on the studying the host-pathogen interactions. It is one of the papers that I am more proud of. Working with you was just so easy and hiccup-free and we wrote a fantastic manuscript together. I still remember, one of the reviewer called our manuscript and supplementary a gold-mine.

I am very very happy that I got to meet you and work with you, **Steinar**. I learnt so many things from you. You started every single meeting by recalling what we had achieved in the previous meeting or project and what an excellent job we had done and how happy you were about it. I will definitely try to replicate your management style when I find myself in a supervisory or leadership position. But I have to say having these long and fruitful scientific conversations with you has been one of the highlights of my PhD. I have always loved a good discussion all my life, so thank you Steinar for the great collaboration, introducing people with the correct expertise and all the great discussions.

I would also like to thank **Anna**, **Mattias**, and **Ole** for all the strong collaborations and fun conversations and dinners during all the PerMIT and PerAID meeting that we met at. Particularly, I would like to thank **Knut** for always engaging with me and thinking highly of my contributions. Already back in 2019, when we first met in Olhao, you walked up to me to tell me that you had seen the analysis I had done, and what an amazing job I had done. Thank you for always making me feel so comfortable.

Oddvar, you have been solely responsible for 100% of my energy drink consumption during this PhD (in order to read all the amazing papers you have sent me). Since Steinar introduced you and included you in some of my projects, you have managed to turn my work from "here's some analysis I did" to "here's an excellent result worth publishing". Receiving the emails from you with an attachment called "Sanjee.docx" has been one of the high points of this PhD. Thank you for your expert contributions, it brought me a lot of happiness.

I would also like to thank all the of co-authors and collaborators from the manuscripts included (and also not included in this thesis), including, but not limited to **Zoe**, **Xi-**

aodan, **Ariette**, **Ingrid**, and so many others. Thank you all for your excellent contributions in this journey.

When I was doing my master's thesis, I got invited to this dinner with some other people from SSB at the house of some guy named **Rik**. While playing some games there, it turned out that we shared more than a few things in common. Later we shared an office together as I started my Phd. Slowly this has turned into a great friendship outside the office. I definitely asked you some really stupid questions at the start of my PhD, but you always helped with no judgments ever. I am grateful for all the fun times we have had together and the ones we ~~will~~ ought to have in the future. However, I can not thank you alone here, I have to also thank **Tjaša** as without her and Shruti planning all the events, we wouldn't have met half the times we did. Even when we met without them, it was because they were going for a movie and we had nothing to do lol. Tjaša you are always so full of energy and enthusiasm and always planning the next thing to do. Never change.

Sasja, I do not remember when we first met, especially because I have such a good memory *sarcasm*, but then you became our neighbor and we had quite a few good times together. But imho, we connected more strongly on that one random drive home, when our discussion turned to neurodivergence. You seemed to just understand exactly what I was saying every time and I opened up so fast and hopefully vice versa. Also, thank you for dragging me out of the house around my thesis submission and the supportive feedback on the thesis cover. I am truly happy to have found a friend in you.

Wasin, we shared an office briefly when I was doing my master's thesis and you taught me how to make a for loop in python. But we only really got to know each other properly during the end stage of your PhD. Even then in a short time, we had a lot of fun together. We had a lot of interesting conversations during lunch and after work. I'm thank full for all the fun times playing board games, dinners, coffee tasting, beer tasting, hot pots and more with you and **Som**. Your stress free attitude and "sure" has been thoroughly missed in SSB. I am sorry, we couldn't make it to Thailand for your marriage, but we are planning to make a trip there after Shruti's defense.

Nhung, you have been a wonderful friend and it's been a lot of fun for years. Since I joined SSB, you have always invited me everywhere for all sorts of activities and to eat Vietnamese spring rolls. We also had a lot of fun being paranymphs for Rik. Honestly, there have been too many fun times to recall all of them. I am happy to have played my tiny role in your story of meeting **Christos** lol. You are honestly a very strong person and your ability to never quit is impressive, however, I'm still waiting for you to start your own company and hire me! :D. I hope we have many more fun times in future with you and Christos. Christos, your blaze attitude and no regard for customs or authority always makes me chuckle.

Sara B, we shared so much of this journey together. We shared an office during our master's thesis. We finished our master's a few months apart, got offered a PhD around the same time and now here we are defending our PhDs a few months apart. You have always made me feel welcome, comfortable, and always cheered on any good work that I did. Even though, we had so many things going on and were under stress, we had quite some fun making Wasin's video. I wish you all the best for your future career move, I'm sure you will take some stress, but in the end produce excellent results.

Sabine, we shared an office twice and you fed me far far far too many liquorice. You are always so cheerful and fun to talk with. Remember, you have promised to speak Nederlands with me so that I finally learn. **Willemijn**, you are the best, perfect and the most reliable secretary SSB could ever wish for. So many times people from other departments have asked me, how I did something and my answer has always been, "I don't know, Willemijn did it!" Even close to my thesis submission, you handled the absurd situation of my email renewal without it ever being a stress to me. So many things have been a breeze and non-issue because of you. Additionally as it turns out, you are also one of the best laser tag players! **Henk** and **William**, it has been an absolute pleasure to share the office with both of you. We have had many interesting discussions on infrastructure, geography, differences with the US, politics, careers and so many more topics. Henk, we more often than not took the opposite stance in the arguments, but that's how I have always liked the discussions to be. I wonder, how any work ever got done in our office. I am also thankful to you Henk for reading parts of my thesis and giving useful comments on them. Maybe we should plan some evening of drinks in a bar for our discussions.

I also want to thank **Architha, Bart, Benoit, Brett, Christos, Claudia, Cristina, Enrique, Erika, Jasper, Jenny, Linda, Lyon, Marco, Maria M, Niels, Nirupama, Peter, Pieter, Rob, Ruben, Silvia, Sonja,** and others who are and were part of SSB (including bachelor and master thesis students) and made the days in the office and lunch conversations fun and interesting. I also would like to thank **Yasmijn, Oriol, Chen, Robert, and Merijn**. Supervising you was very rewarding, and I learnt new skills supervising each one of you. I hope you were able to take away something positive as well.

Finally, I want to thank my family members for their never-ending, unconditional support and cheer leading every step of the way. My sister **Amruta**, you have always taken the role of elder sister way too personally. Nobody dare say anything negative about me, you will always be there defending my choices, way of life, and opinions. Thank you for all the great memories, 3am dances & games, buying me my first cell phone, and just being there for me every single time. **Amit**, you wanted to buy me my first ticket to the Netherlands as gesture of an opportunity like you had going to Germany, but instead I asked you to buy me a big storage drive. That storage drive now holds monthly backups of every single thing I have done in this PhD. So, in some ways, given my job, you did complete your gesture. **Aaniya**, it is has been a privilege getting to know you. You are always playful, joyful and happy. Yet, never afraid to ask the truly difficult questions where no amount of philosophical frolicking can stop the next inevitable why. And last, but not the least, my aai, **Simantini** and my daddy, **Rajendra**, Thank you for providing me with a safe space, the appreciation of all my work, and everything, and that is just for the years when this book was written. Thank you for everything; for the childcrafts, mechanic games that said 17+ when I was like 7, teaching me to use the internet (it was yahoo search back then, google didn't even exist), telling me the story of evolution by natural selection... I don't know, the list is endless and who can measure the effect your actions have had on this thesis. So, I would just end by saying Thank you.

About the Cover

The cover of this thesis symbolises the convergence of two distinct scientific domains investigating Necrotising Soft Tissue Infections (NSTI). The dark blue represents the biological realm, while the dark green signifies the computational dimension. In the center, a humanoid figure serves as a bridge between these two realms, embodying a network of interacting entities. The reader is left to contemplate & speculate whether the computer simulates the humanoid, and by extension, the biological world, or if the humanoid is an integral part of the biological realm and modeled on the computer.

As the title suggests, "Peeking Under the Skin," the cover delves beyond the visible surface, typically occupied by the humanoid in real life, to explore the hidden dynamics of these diverse realms. An image of the Thrombomodulin protein is featured within a magnifying glass peeking under the skin, but it also transcends the confines of the magnifying glass, engaging with the biological world on a different scale. The differential effect on Thrombomodulin concentration has been one of the important findings from this thesis. The computational side of the cover also symbolises my involvement and contributions to the topic, all accomplished through computational work. Moreover, the backdrop behind the computer is an ASCII binary code representing text in computers. The ASCII binary code, if translated, conveys the message "Take the risk of thinking for yourself", a quote that is very meaningful to me. Thus, portraying the human element (in this case myself) running the simulations on the computer.

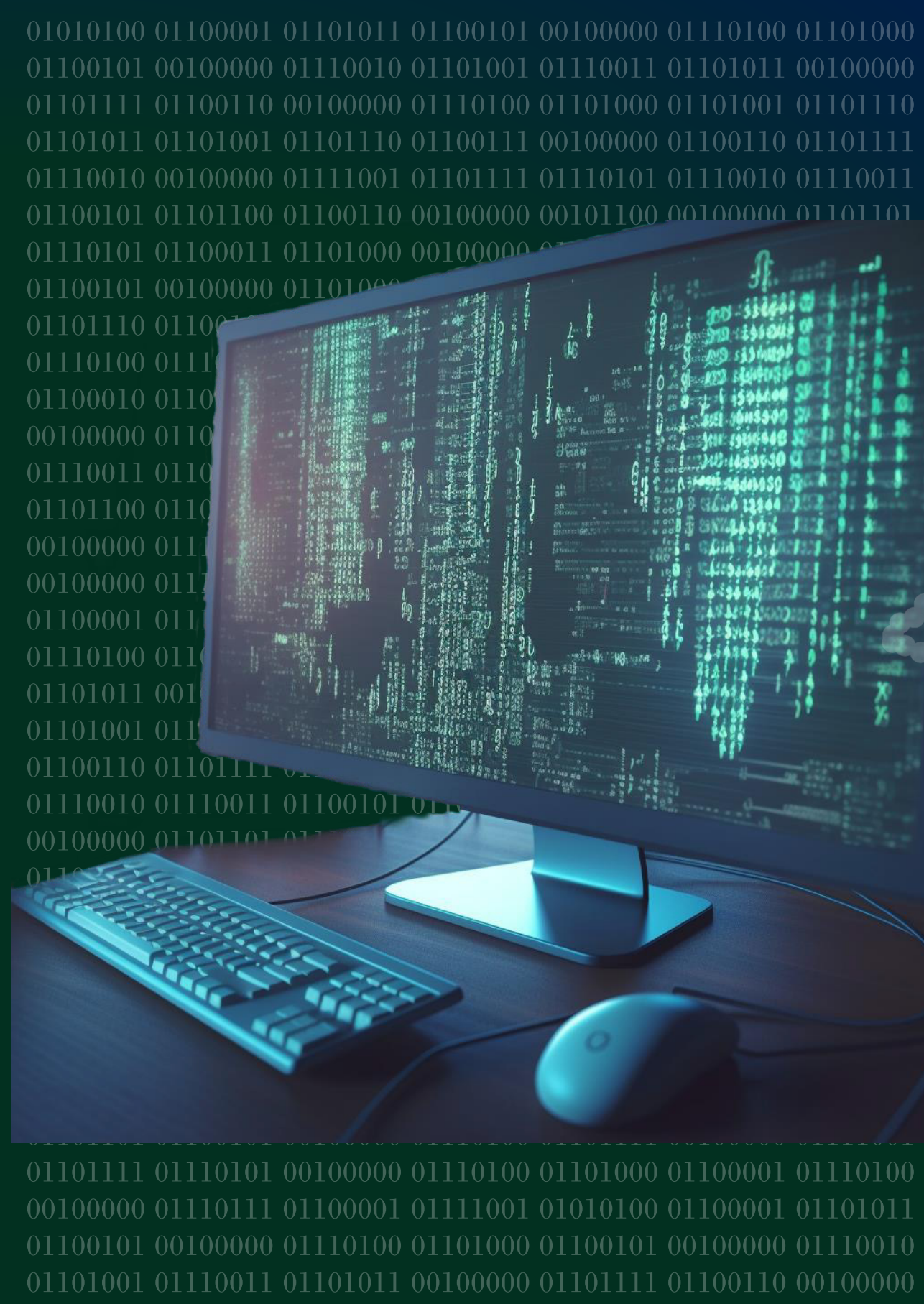
On a philosophical note, the cover underscores the seamless collaboration between computational and biological studies. However, it also emphasises the human element that is often overlooked — the individual impacted by or benefiting from these endeavors. It serves as a reminder that these studies exist within a broader ethical and moral context.

Midjourney (AI) was used to generate and blend images in order to produce some of the images seen in the cover and artworks before each chapter.

The research described in this thesis was financially supported by the Netherlands Organisation for Health Research and Development (ZonMW) through the PerMIT project with project number 456008002, NordForsk through the PerAID project with project number 90645, the European commission (FP-7-Health) through the INFECT project with project number 305340, and Wageningen Data Competence Centre (WDCC) through the DS/AI fellowship with project number 2200000200

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Printed by ProefschriftMaken || www.proefschriftmaken.nl


01010100 01100001 01101011 01100101 00100000 01110100 01101000
01100101 00100000 01110010 01101001 01110011 01101011 00100000
01101111 01100110 00100000 01110100 01101000 01101001 01101110
01101011 01101001 01101110 01100111 00100000 01100110 01101111
01110010 00100000 01111001 01101111 01110101 01110010 01110011
01100101 01101100 01100110 00100000 00101100 00100000 01101101
01110101 01100011 01101000 00100000
01100101 00100000 01101000
01101110 01100
01110100 0111
01100010 0110
00100000 0110
01110011 0110
01101100 0110
00100000 011
00100000 011
01100001 011
01110100 011
01101011 001
01101001 011
01100110 01101111 01
01110010 01110011 01100101 011
00100000 01101101 011
011
01101111 01110101 00100000 01110100 01101000 01100001 01110100
00100000 01110111 01100001 01111001 01010100 01100001 01101011
01100101 00100000 01110100 01101000 01100101 00100000 01110010
01101001 01110011 01101011 00100000 01101111 01100110 00100000