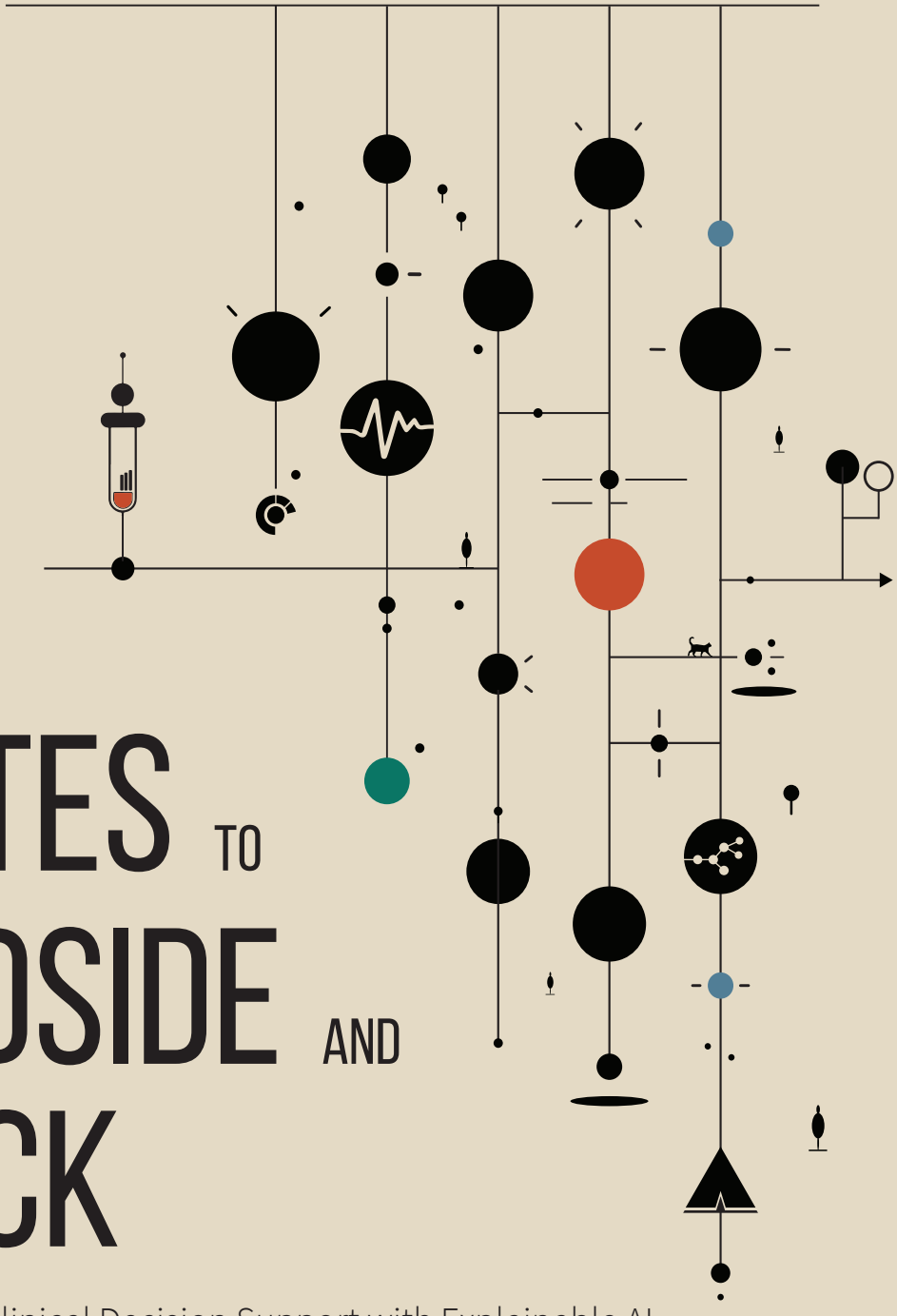
SONJA KATZ



FROM BYTES TO BEDSIDE AND BACK

Enhancing Clinical Decision Support with Explainable AI

Propositions

1. Decision-support systems are essential in realising precision medicine in clinical practice.

   (this thesis)

2. Delineating the decision-making of AI models holds the key to enhancing our understanding of health.

   (this thesis)

3. Modern science is ruled by the latest methodological trends.

4. An accurate but opaque predictive AI model is better than doctors' subjective decisions.

5. Precision medicine will increase social inequality.

6. Scientific reasoning should be taught in schools to counteract fake news.

Propositions belonging to the thesis, entitled

From Bytes to Bedside and Back:
Enhancing Clinical Decision Support with Explainable AI

Sonja Katz
Wageningen, 20 September 2024

From Bytes to Bedside and Back: Enhancing Clinical Decision Support with Explainable AI

Sonja Katz

Thesis committee

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

Co-promotors
Dr E. Saccenti
Associate professor at the Laboratory of Systems and Synthetic Biology
Wageningen University & Research

Dr G.V. Roshchupkin
Assistant professor and Computational Population Biology group leader
Erasmus MC Medical Center, Rotterdam

Other members
Prof. Dr A. Fensel, Wageningen University & Research
Prof. Dr J. Camacho, University of Granada, Spain
Prof. Dr F. Rivadeneira, Erasmus Medical Center, Rotterdam
Dr W.J. Santos Silva, Utrecht University

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

From Bytes to Bedside and Back: Enhancing Clinical Decision Support with Explainable AI

Sonja Katz

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr C. Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 20 September 2024
at 10.30 a.m. in the Omnia Auditorium.

Sonja Katz
From Bytes to Bedside and Back:
Enhancing Clinical Decision Support with Explainable AI,
305 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2024)
With references, with summary in English

DOI: https://doi.org/10.18174/662462

The Financial support from the TranSYS International Training Network from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No. 860895 is gratefully acknowledged.

Contents

Chapter 1

Introduction: on Precision Medicine, Artificial Intelligence, and their Symbiosis

> It is not my aim to surprise or shock you – but the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be co-extensive with the range to which the human mind has been applied.
>
> *Herbert Simon, 1957*

Parts of this chapter will be prepared for publication

Throughout history, medicine has undergone multiple transformations in its approach to personalisation. During times in which medicine was still guided by mysticism, religion, and patient's faith in the healer - the efficacy of which would today be attributed to the placebo-effect [1, 2] - treatments could be regarded as highly individualised due to the sheer lack of knowledge available to practitioners.
In the 1800s, a time still governed by the belief that diseases were caused by "miasmas" - noxious forms of air [3] - sceptical thinkers began to challenge these ideas. Notably, without understanding the exact mechanisms, physician John Snow traced the source of a cholera outbreak in a London district solely through careful observation of facts [4]. His studies, now regarded as the founding event of modern epidemiology, laid the cornerstone for redefining centuries-old medical paradigms and introduced a new concept: evidence-based medicine. With its focus on clinical trials including large-scale studies, randomisation, and double-blind procedures, evidence-based medicine allows clinical decision-making to be more structured, objective, and tied to external clinical evidence from systematic research, as opposed to the beliefs of experts [5]. However, the rise of evidence-based medicine came at the expense of personalised treatment. With its focus on identifying generalised standards in large cohorts of patients, evidence-based policies leave little room for patients not fitting the respective norm [1]. With the non-responder rates in clinical trials not seldom exceeding 50% [6], the question of what happens to these patients remains unanswered although "inter-patient variability in response to drug therapy is the rule, not the exception, for almost all medications" [1, 7].
Accelerated by the achievement of big scientific milestones such as the completion of the Human Genome Project [8], the 21st century saw the acknowledgement of patient diversity gaining more and more momentum. Critical voices regarding the principles of evidence-based medicine, especially applying population means to individual patients, became louder, a concern that Kravitz *et al.* fittingly termed "trouble with the averages" [9]. This momentum shaped the concept that would come to be known as precision medicine.

1.1 What is Precision Medicine?

> Patient heterogeneity was a nuisance for evidence-based medicine but a blessing for precision medicine.
>
> *M. R. Kosorok et al., 2019* [10]

Precision medicine, or personalised medicine, describes a scientific approach wherein

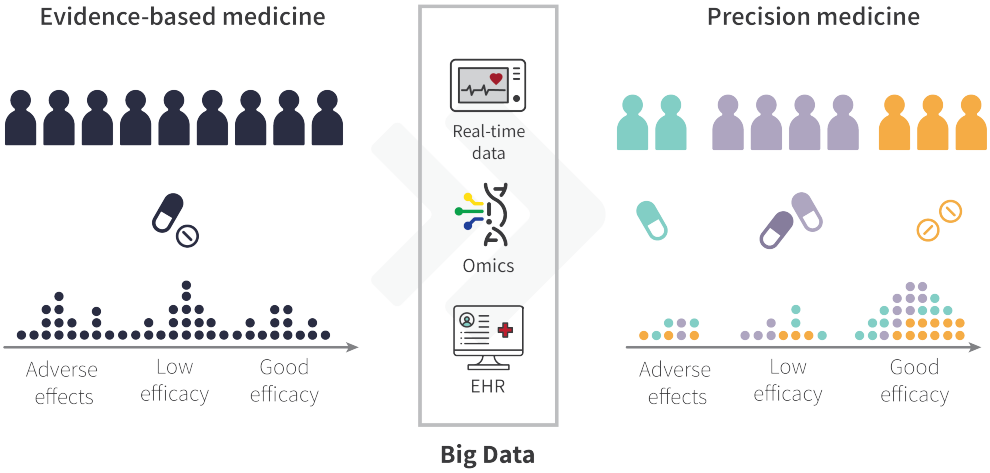


Figure 1.1: **Traditional evidence-based medicine *versus* precision medicine.** Adapted from [11].

patient heterogeneity is leveraged through data-driven approaches [12, 13]. Although the term is not easily summarised, the European Commission defined that precision medicine "refers to a medical model using characterisation of individuals' phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention" [14].
The objectives of precision medicine therefore encompass several distinct aspects [12, 15–17]:

(I) Personalised pharmacogenomics

   i. improve the medication effectiveness for patients

   ii. reduce the risk of adverse treatment effects by avoiding therapies showing no clear positive effect on the disease, while at the same time exhibiting (partially unavoidable) negative side effects

(II) Diagnosis and prevention

   i. establish early disease diagnosis and prevention through molecular and non-molecular biomarkers

(III) Disease management

   i. improved disease management with the help of wearable sensors and mobile health applications

(IV) Healthcare economics

i. lower healthcare costs as a consequence of effective use of therapies

ii. smarter design of clinical trials due to selection of likely responder at baseline

1.1.1 Do we need Precision Medicine?

It comes as no surprise that the 21st century's prevailing trend toward individualism is reflected in our expectations of healthcare systems. However, why should we care about precision medicine? Do we even really need it?
I present here two scenarios where evidence-based medicine reaches its boundaries, and where patients could potentially benefit from the application of precision medicine principles.

Necrotising Soft Tissue Infections (NSTI)

> *Frankie became unwell on Monday 8th April 2013 and he was only a 1 year of age. He had flu like symptoms. (...) In the late morning of Tuesday 9th April, Frankie appeared to become very unwell with similar pustules on his face and a red pinprick rash all over his arms, chest and back. (...) Frankie looked totally delirious. He could barely move. I noticed that the skin on Frankie's back was now red in parts. (...) The doctors started putting what she said were antibiotics into Frankie's cannula and pushing large syringes of saline into him one after the other. This was when I became worried something was really, seriously wrong. (...) Frankie was in surgery for almost 10 h (...). We were told that Frankie's body had been extensively damaged by the infection. The surgeon had removed all the layers of skin and fascia from almost his entire back, left side and quite a lot of his left thigh and muscle had been removed. We were told that they had not been able to remove all the infection because Frankie was too weak for any further surgery and that he was been given strong antibiotics intravenously. The consultant told us that they had done everything they could at the point and that it was now up to Frankie's body to fight the infection with the help of some broad-spectrum antibiotics. We were asked to get everyone we needed up to the hospital because they did not believe our baby was going to survive the night.* - Lucy Dove, Frankie's mother [18]

Frankie, 12 years old today, is a survivor of the devastating infectious disease termed necrotising soft tissue infection (NSTI) [18]. NSTI are caused by bacteria and may affect

any layer of the soft tissue, and thus can appear on any part of the body. Presentations of the disease can therefore vary considerably [18] but commonly include symptoms perceived as "mild", such as influenza-like symptoms (fever, nausea), as well as swelling, redness of the skin, and pain in the affected area [19, 20]. However, the frequent absence of fulminant symptoms may not indicate the lack of seriousness of NSTI's fatality: patients frequently suffer from septic shock, a life-threatening condition (28-50%), with mortalities ranging between 10-29% [19, 21–25]. The lucky survivors have to deal with a significantly reduced quality of life due to extensive scarring, amputations, and psycho-social burdens [26–29].

But why does NSTI present such a challenge to our modern healthcare system? As evident from the case report, NSTI progresses rapidly - if left untreated, it can be fatal within days [18]. Paired with the vague initial symptoms that cause patients to seek medical help only after several days of experiencing symptoms at home [30], NSTI is frequently diagnosed when it is already too late. Adding to this, the disease is very uncommon, affecting only 1.8 - 15 per 100,000 inhabitants per year [31], and is thus frequently underdiagnosed. Once successfully recognised, clinical decisions do not get easier as NSTI presents itself as highly heterogeneous with regards to age, sex, the presence of comorbidities, and disease-causing micro-organisms [32], with its pathophysiology still incompletely understood [33].

So how can modern precision medicine concepts improve NSTI care? Firstly, the establishment of novel biomarkers could help in diagnosing patients earlier and more reliably. Palma Medina *et al.* report thrombomodulin to be a unique and powerful biomarker for the detection of NSTI. Additionally, it is possible to differentiate between specific clinical phenotypes of NSTI based on the inflammatory profiles of patients, which has direct implications regarding the choices of therapeutic intervention [34]. The observed differences in inflammatory response are likely driven by pathogen-specific underlying mechanisms, which underlines the necessity to further explore host-pathogen interactions. In an attempt to anticipate patient's individual responses to infections, Jahagirdar *et al.* have studied NSTI type-specific responses in the patients [35]. They found that the degree of dysregulation of the host response to infection is directly linked to the severity and outcome of NSTI. Concerning the patient's phenotype, they were able to postulate differing modes of entry and immune evasion for different bacteria, which can be regarded as the first step in identifying novel targets for more personalised interventions.

Despite these studies presenting early endeavours in applying precision medicine to NSTI, the compelling results presented underscore the significant potential that precision medicine holds for advancing the diagnosis, prognosis, and therapy of NSTI.

COVID-19

At the time of writing this thesis, which is January 2024, it feels superfluous to introduce COVID-19. However, for the sake of future (or oblivious) readers, here I offer a brief introduction to COVID-19 and how patients could have benefited from healthcare systems employing modern precision medicine concepts.

The coronavirus disease 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) was first reported in December 2019 and already declared a "public health emergency of international concern" by the end of January 2020 [36]. A total of 774.075.242 COVID-19 cases were reported to the WHO as of January 2024, accompanied by more than 7.000.000 deaths [37]. These horrifying numbers place COVID-19 on the fifth rank of the list of the deadliest epidemics and pandemics in history [38].

An infection with the SARS-CoV-2 virus can show diverse manifestations ranging from asymptomatic infections to severe, life-threatening conditions. Most commonly the symptoms include fever, a sore throat, and fatigue, with recovery times ranging from two weeks for mild infections to prolonged ICU stays for severe cases. Between 5-50% of patients suffer from persisting long-term consequences, including but not limited to fatigue, shortness of breath, sleep problems, and anxiety [39, 40]. Initial COVID-19 treatments consisted of antiviral and anti-inflammatory drugs - originally developed to treat malaria, HIV, and Ebola - which showed variable efficacy and frequent adverse reactions. In an unprecedented joint global effort, the first vaccine was introduced only 8 months after the first COVID-19 case was reported [41]. Mass vaccination campaigns followed, with more than 13.59 billion doses administered worldwide as of today [37]. However, the developed vaccine was designed to target whole communities, independent of age, comorbidities, sex, or geographical region. To no surprise, this led to dramatic differences in preventive efficacy and adverse side effects of vaccinated [41].

The heterogeneity of COVID-19 disease courses, treatment response, and vaccination efficacy is attributable to a diverse set of factors, such as virus variants as well as human immune response determinants, genetic makeup, age, sex, and ethnicity [42]. While this heterogeneity poses a significant hurdle for evidence-based medicine, precision medicine can help with elucidating individual susceptibility to the infection as well as inter-individual variability in clinical course, prognosis, and response to treatment. By following the basic principles of precision vaccinology, for instance, we can develop improved vaccines delivering tailored immunisation for vulnerable populations with distinct immunity [43]. To control the variable drug responses and limit adverse outcomes following therapies with antiviral drugs, stratification of patients according to their inflammatory profiles has

been able to identify distinct sub-phenotypes in patients with similar clinical features. These sup-phenotypes hold substantial clinical significance. For instance, convalescent plasma therapy, a therapeutic approach utilising blood from individuals who have recovered, demonstrated beneficial effects in one subgroup (with a 3% decrease in mortality) but appeared harmful in another subgroup (with an increase in mortality of up to 11%) [41, 44] Precision medicine also aims at redefining aspects such as disease management and healthcare economics. In hindsight, the COVID-19 pandemic could have greatly benefited from established digital health solutions, as a lack of data in the initial stages of the outbreak led to its rapid spread across the world [45]. Smart healthcare solutions, such as pandemic monitoring apps, could have empowered policy-makers to swiftly respond to emerging developments. Additionally, mobile health apps and telemedicine would have enabled citizens to avoid unnecessary visits to healthcare institutions, thereby alleviating the strain on hospital resources [46].

1.1.2 Integrating Precision Medicine into Clinical Practice: Clinical Decision Support Systems

While the potential benefits of precision medicine are substantial, there does not exist a uniform concept on how to integrate it into clinical routines. While there are many ideas and concepts, they can easily seem abstract and lack applicability. So might the specific implementation of holistic and data-driven precision medicine applications in clinical routines look like?

For inspiration, let's look at how visionaries of our time imagined the integration of advanced technology in everyday lives: Science Fiction writers. The Emergency Medical Holographic program (EMH) created in the fictional *Star Trek* future is a prototype designed to provide medical assistance during emergencies. Portrayed as a human-looking hologram it is programmed with over five million possible treatments from around 2000 medical references. Equipped with adaptive programs to learn over time and fifty million "gigaquads" of computer memory, it supplements humanoid medical staff [47]. In combination with another standard equipment found on Federation starships, the medical tricoder, which is a high-resolution, hand-held scanner used to check all vital organ functions and equipped with an extensive data bank even enabling the treatment of other life forms, Starfleet officers are well-equipped to deal with all cosmic eventualities [48].

While EMH and tricorders may seem preposterous - and will only be developed in the early 2370s - they encompass the core idea of a precision medicine tool: a piece of equipment or software clinicians can (i) interact with via an interface, capable of (ii) evaluating patient information and delivering patient-specific recommendations based on (iii) an

extensive knowledge base, which are ultimately presented to the clinician for consideration and decision-making. Fortunately, such a tool is not limited to Science Fiction novels anymore; rather, it is rapidly gaining popularity and is commonly referred to as a Clinical Decision Support System (CDSS) [49]. A CDSS commonly operates across multiple levels, incorporating data management, modelling and prediction, and visualisation into a single platform. At its core, CDSS rely on inference machines, which commonly are predictive models employing machine learning or deep learning methods to identify diseases, disease stages, and patient-specific patterns [50]. In clinical practice, CDSS serve multiple valuable functions, such as giving recommendations and suggestions, creating automated alerts and reminders in the ICU, or reducing clinical error through e.g. complicated calculations of treatment dosing [51]. One of the biggest advantages the use of CDSS entails is having the possibility to easily take into account information from multiple different sources and modalities, abrogating the need for resource-intensive manual comparison of evidence from e.g. different medical guidelines. Another benefit lies in their circular, iterative design. By constructing predictive models atop a knowledge database system, the addition of new patients or updates to available information can occur in an automated fashion. Consequently, this enhances the accuracy of recommendations over time, as the system continuously learns and adapts with each piece of added information [49].

1.1.3 The Data Powering Precision Medicine

Precision medicine re-imagines the functioning of our healthcare approaches and systems by placing the individual patient at its core. But how does it aim at realising this ambitious concept? The key to success lies in the collection of evermore information about patients, diseases, and treatments. The more information we possess, so the idea, the better we can make decisions that deliver the best outcomes. This paradigm led to the collection of immense amounts of data in health sciences. Accelerated by the broad availability of technological innovations such as next-generation sequencing (NGS), patients can now not only be assessed using their demographics and a handful of clinical parameters, but are characterised by a wealth of genomic, imaging, and real-time data [52]. However, not only the amount of data is tied to its clinical usefulness, but also its quality. A common saying in machine learning nicely summarises this: "garbage in, garbage out". Missing data points, convoluted and inconsistent representations, subjectivity, or error-prone measurements all severely diminish the quality of data, thereby limiting the representativeness of models trained on them [53].

The overload of information available - also referred to as 'Big Data' (Figure 1.1) - is what provides precision medicine the needed depth and what distinguishes it from traditional

evidence-based medicine [15, 54].
Some examples of data powering precision medicine include:

- **Electronic health records (EHR)**, which are the sum of computerised medical records for patients. This includes demographics, family and medical history, medication and allergies, immunisation status, laboratory test results, vital signs, and radiology images [55].
- **Omics data**, which refers to comprehensive, high-throughput biological datasets. With the establishment of efficient technologies like NGS, we are now able to observe biological events associated with specific diseases at unprecedented resolution. Before clinical symptoms manifest several aberrations in biological processes happen - whether it be genetic mutations, epigenetic alterations, or post-translational defects. Each type of omics measurement therefore reveals an aspect of cell complexity. Although it is impossible to curate a comprehensive list of all omics types due to the speed at which technologies are developing, the most important areas include genomics, epigenomics, transcriptomics, proteomics, and metabolomics [56].
- **Real-time (RT) data**, which includes the continuous collection, monitoring, and analysis of patient-related information. With the advanced capabilities of personal devices like smartphones and smartwatches, which can count steps, monitor heart rate, and assess sleep quality, there is a growing interest in leveraging real-time data for the early diagnosis of cardiovascular diseases [57, 58], Alzheimer's [59], and for monitoring the health and safety of the elderly [60].

The mass of data we are now able to provide for each individual has not only revealed a new perspective on the intricate way biological systems function but has also re-defined our requirement for tools that can be applied to analyse this wealth of heterogeneous information. Many diseases that pose a challenge for today's healthcare system such as NSTI, COVID-19, cancer, diabetes, metabolic syndrome, and neurological, or immunological disorders are complex, affecting several biological sub-systems [61]. It is close to impossible to e.g. find a single genomic biomarker stratifying a lung cancer population in drug responders and non-responders [62], or identify an easily measurable clinical parameter for early diagnosis of metabolic-associated fatty liver disease [63]. Also, when aiming to make causal inferences from observational data, one needs to consider - and potentially correct for - the presence of systematic biases, such as confounding factors. Confounding factors are variables associated with both, the exposure and outcome of interest. This "mixing of factors" can obscure a true association or, more frequently, simulate an association, leading to false conclusions [64].

In order to assist patients, we must seek computational tools capable of interpreting complex, interconnected, biased, and heterogeneous data across multiple dimensions. Artificial intelligence stands out as a potential answer to our question.

1.2 From Hype to Reality: Artificial Intelligence enabling Precision Medicine

An unprecedented wealth of data also requires an unprecedented innovation in tools analysing respective data. It can therefore be argued that the personalised health care revolution we currently find ourselves in is only made possible by these tools, of which artificial intelligence is at the epicentre.

1.2.1 What is Artificial Intelligence?

Artificial intelligence (AI) encompasses a range of methods that empower computers to imitate human actions and perform complex tasks autonomously or with minimal human intervention, often surpassing human decision-making capabilities [65]. Machine learning (ML) and deep learning (DL) are both subcategories of AI; their hierarchy is illustrated in Figure (Figure 1.2a). Therein, deep learning is a subset of machine learning that relies on artificial neural networks, a structure inspired by the human brain, to process and analyse information.
While early AI research focused on hard-coded rule sets that can be parsed to infer logical assumptions about the data, modern approaches focus on developing algorithms that can learn from data it was not explicitly programmed for. "Learning" in this sense means that, given a problem, an algorithm's performance measures (accuracy, error rate) improve with experience - the more examples of an instance a model has seen, the better can abstract the problem [66]. The aforementioned problems an algorithm may encounter can be grouped into one of the following categories [67] (Figure 1.2b):

(I) **Supervised:**

(i) **Classification:** problems in which the output can only be one of a fixed number of classes e.g. predicting whether a patient died or not (binary), which cancer subtype is present (multi-class).

(ii) **Regression:** problems in which the target output is a continuous value e.g estimating how many days will a patient stay in the intensive care unit (ICU).

(II) **Reinforcement Learning:** tasks the model to learn to make decisions based on past experiences from trial and error interactions to achieve a clearly defined goal e.g. deduce treatment policies for septic patients to maximise their survival [68].

(III) **Unsupervised:**

(i) **Anomaly Detection:** problems in which patterns are analysed and the task is to detect examples that do not follow the general pattern, e.g. finding abnormalities in medical images [69].

(ii) **Dimensionality Reduction:** process of simplifying a dataset by removing noisy and irrelevant data while preserving as much information as possible; e.g. training a model only on selected variables correlated with the outcome of interest.

(iii) **Clustering:** learns structure within the data and attempts to group similar samples together, ultimately forming e.g. clusters of related patients.

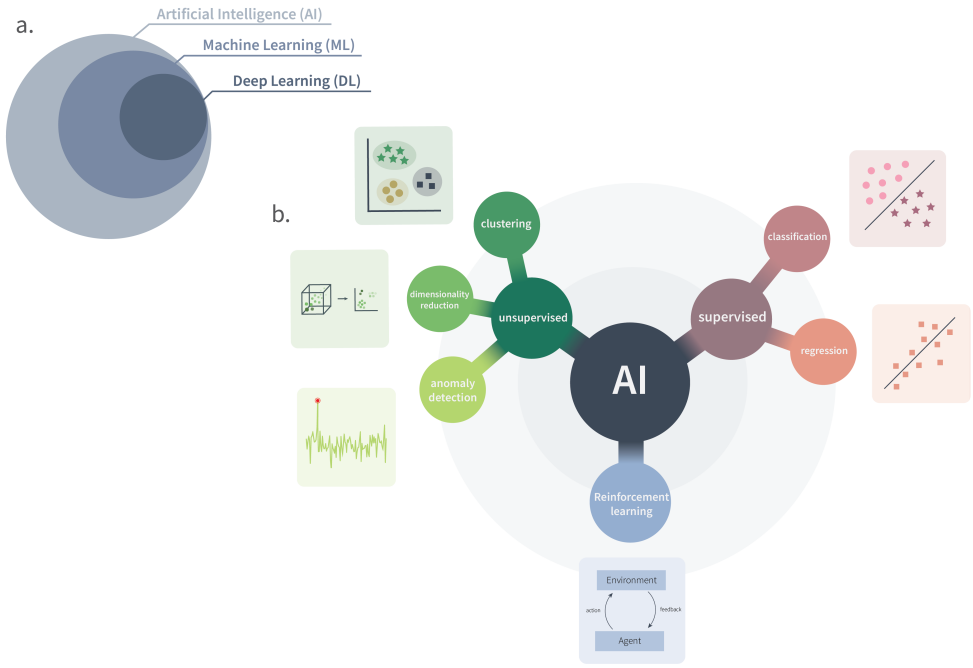


Figure 1.2: **Artificial Intelligence (AI) and its categories**. (a) Hierarchy of AI. (b) Categories of problems AI can be trained on: supervised (classification, regression), reinforcement learning, and unsupervised (anomaly detection, dimensionality reduction, clustering).

The following section offers an overview of how machine learning and deep learning can

be applied to biomedical data. I will elucidate the distinctions between the two and why precision medicine cannot thrive solely with either of these approaches.

1.2.2 Selecting the Right Tool for the Job: Machine Learning or Deep Learning?

There are an overwhelming abundance of artificial intelligence algorithms, each with its own application area, strengths, and weaknesses. A key decision for data scientists at the start of each new project is the choice of which algorithm is most fitting to solve the question at hand. Although there exist a number of factors influencing this decision, two major things should be considered: firstly, in which category of the ones outlined above falls the problem (classification, regression, clustering) and secondly, what the nature of the data is that the analysis will be based upon. While it is often straightforward to determine the desired outcome, understanding the nature of the data that will be used is more complex. Researchers need to not only consider the number of variables measured, the type of the variable (continuous, discrete, text), and the number of samples (or patients) included but also the relationship between variables or between variables and measured outcome. Being aware of the structure of the data is of central importance, as algorithms make different assumptions about the data, resulting in varying capacities to abstract the data. A well-versed data scientist is aware of this and selects the right tool for the job.
To introduce various models featured in this thesis and elucidate the considerations going into algorithm selection, I will showcase three progressively complex medical examples or problems. By looking at a regression, classification, and clustering scenario, I aim to explore the limits of machine learning and demonstrate that, although powerful, the application of deep learning algorithms does not come without significant challenges. It is important to note, however, that the scenarios presented are greatly simplified and designed to illustrate the concept of how to apply machine and deep learning concepts to medical questions. An overview of the relationship between the complexity of the data, model performance, and indication of which areas the outlined medical problems reside in can be found in Figure 1.3.

(I) Simple Problem: Estimating Cholesterol Levels of Patients from Clinical Data

Problem summary:

- **Task**: predict the cholesterol levels of patients
- **Number of features:** low (<10 features)

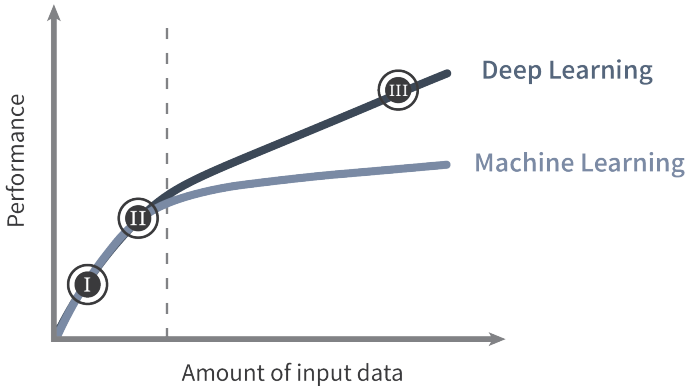


Figure 1.3: **Relationship between input data and AI model performance**. x-axis denotes the amount (and complexity) of input data, whereas y-axis denotes the achievable performance of models. The dashed line denotes the inflection point where the performance of deep learning and machine learning models diverges with increasing data volume. (I), (II), and (III) indicate the positions of exemplary medical examples/problems categorised as simple, intermediate, and difficult, respectively, as discussed in this section.

- **Complexity of feature interactions:** low (linear interactions)

In the first presented scenario, the medical problem we encounter deals with estimating the cholesterol levels of patients. A sufficiently large number of patients was recruited (>1000 patients), and a few demographic and clinical parameters were assessed, including age, blood pressure, body mass index (BMI), and others. As shown in Figure 1.4, we observe a linear relationship between the target (cholesterol levels) and features (age, blood pressure).
Due to the low number of features and the simple interactions between the dependent and independent variables, we decide to train a linear machine learning model, such as linear regression, for predicting the cholesterol levels of future patients.

Linear regression - and logistic regression, its extension for binary and multiclass outcomes - are amongst the most simple class of algorithms as they model the relationship between a target and one or more features by fitting a linear equation to observed data [70]. Despite their simplicity, logistic regression models are of immense clinical relevance as they form the base of a number of clinical outcome prediction scores, including the Mortality Probability Model (MPM) [71], Acute Physiology and Chronic Health Evaluation (APACHE) [72], Simplified Acute Physiology Score (SAPS) [73]. These clinical scoring systems are used indicate of the risk of death of groups of patients admitted to the intensive care unit (ICU); e.g. APACHE is one of the most widely used scores to obtain a severity of illness score for ICU patients, whereas the SAPS score can be used to directly estimate the

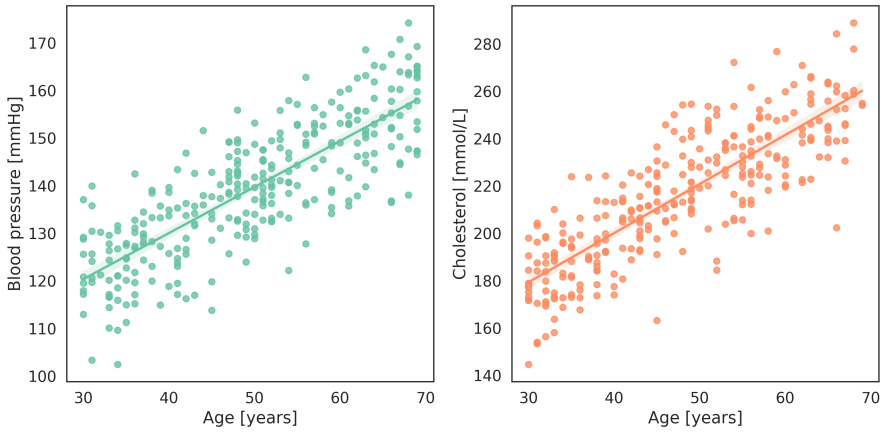


Figure 1.4: **Example of a simple medical correlation.** A linear relationship between blood pressure and age (left), as well as cholesterol levels and age (right) can be observed. Graph generated using simulated data.

risk of mortality in ICU patients.

(II) Intermediate Problem: Predicting the Risk of a Cardiovascular Event from Biomarkers

Problem summary:

- **Task**: predict the probability of a cardiovascular event
- **Number of features:** low (<10 features)
- **Complexity of feature interactions:** complex (non-linear interactions)

In this second scenario, we aim to predict the risk of a cardiovascular event using patients' age and diagnostic test scores. This medical problem, as opposed to the first one presented, is considerably more difficult due to an increase in the complexity of feature interactions. An example of interactions observed in the data is shown in Figure 1.5a; in more detail, we observe a polynomial third-degree interaction between age and the diagnostic score, as well as a sinodial relationship between the target variable (cardiovascular event) and both independent variables (age, diagnostic score). While linear models are among the most widely used methods in medicine and biology, their assumption of linearity constitutes a major limitation, preventing them from modelling the more complex non-linear relationships present within this dataset.
Therefore, we decide to train a Random Forest Classifier (RFC) to predict possible cardiovascular events for unseen patients.

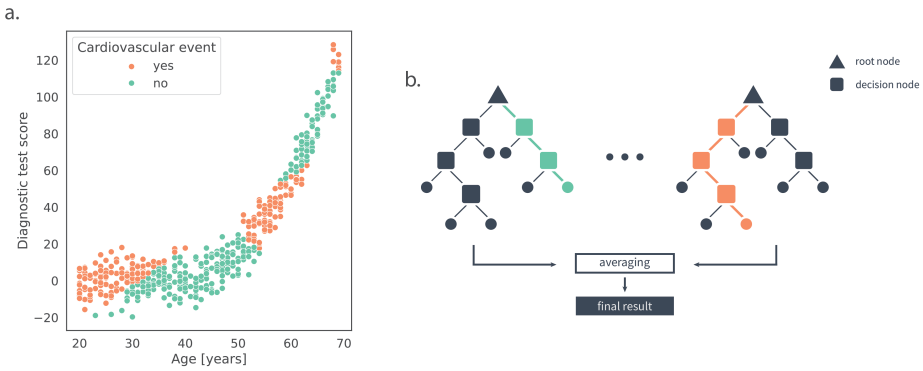


Figure 1.5: **Example of an intermediate medical problem.**(a) Non-linear relationship between a diagnostic test score (y-axis), patients' age (x-axis) and the occurrence of a cardiovascular event (colour). Graph generated using simulated data. (b) Conceptual depiction of Random Forest Classifiers. Two decision trees from the forest are shown, with triangles denote the starting or root node, squares denote decision nodes, and circles denote decision outputs. The decision path taken within each tree is shown in colour. The output of all decision trees in the forest is averaged, yielding a final prediction result.

Random Forest Classifier (RFC) [74] belong to the ensemble machine learning models, as they combine the output of individual smaller classifiers, namely Decision Trees, to one final result (Figure 1.5b). Decision trees possess a flowchart-like structure, consisting of root nodes, which represent the starting point, internal decision nodes, and terminal leaf nodes, holding a class label. The goal is to create a tree that predicts the value of a target variable by learning simple decision rules inferred from the data features at each internal node. Ideally, the learned rules should optimally separate the target classes. Decision Trees and, by extension, Random Forests (RFC) are among the most popular ML methods, due to their capacity to capture non-linear relationships in data and their logical structures, which make them intuitive and easy to interpret [67, 75].

(III) Difficult Problem: Identifying Clinical Sub-Types of Glioblastoma Multiform Based on Their Epigenetic Profiles

Problem summary:

- **Task**: cluster glioblastoma multiform patients based on their DNA methylation data
- **Number of features:** very high (935.000 features)
- **Complexity of feature interactions:** complex (non-linear interactions)

Our last medical problem is motivated by an ongoing challenge in cancer research: the

stratification of glioblastoma multiform (GBM) patients based on their epigenetic profiles [76–78]. GBM is the most common, aggressive, and treatment-resistant primary brain cancer known; the average length of survival for glioblastoma patients is estimated to be only around 8 months [79]. Recent studies have shown the potential to accurately stratify patients using DNA methylation [80, 81]. DNA methylation is typically evaluated using an array that measures up to 935.000 DNA methylation sites per sample. This means, that for us to cluster patients according to their DNA methylation profiles, models need to abstract the relationships between 935.000 features. This is where machine learning algorithms encounter their limits, due to a phenomenon commonly known as the "curse of dimensionality" [65]. The curse of dimensionality refers to the fact that the amount of data required for models to make accurate predictions on previously unseen data, referred to as the generalisability of models, increases drastically with the number of features we include in our analysis. Adequately modelling 935.000 features thus requires recruiting millions of GBM patients - a number which quite frankly does not exist given a 5-year survival rate of only 6.9% [79].
So to identify clinical GBM subtypes, we must expand our repertoire of methods and turn to algorithms which are less affected by the curse of dimensionality. More precisely, we must turn to deep learning.

Artificial Neural Network (NN) are the simplest form of deep learning framework. Drawing inspiration from the information processing in biological neuronal systems, an NN is an interconnected network of individual neurons. Neurons thus serve as the fundamental building blocks of every deep learning framework. They are simple units capable of receiving inputs, performing some processing, and producing an output (Figure 1.6a). While a single neuron may not be powerful on its own, the strength lies in combining multiple neurons to work together, allowing them to theoretically approximate any function. But how does a neural network learn? The learning process of a neural network involves repeatedly adjusting the connection strength between individual neurons. Therein, a sample is passed to the network, which generates a predicted outcome for the respective sample. By calculating the difference between the predicted value and the actual value, the network adjusts its weights to minimise the error, aiming to achieve an output that closely aligns with the ground truth [82].
What makes neural networks a universal tool capable of dealing with diverse medical problems lies in the adaptable arrangement of neurons. For instance, in our effort to cluster GBM patients, we must initially reduce the dimensionality of our dataset as clustering algorithms like K-means [83] are incapable of working with all 935.000 features. We

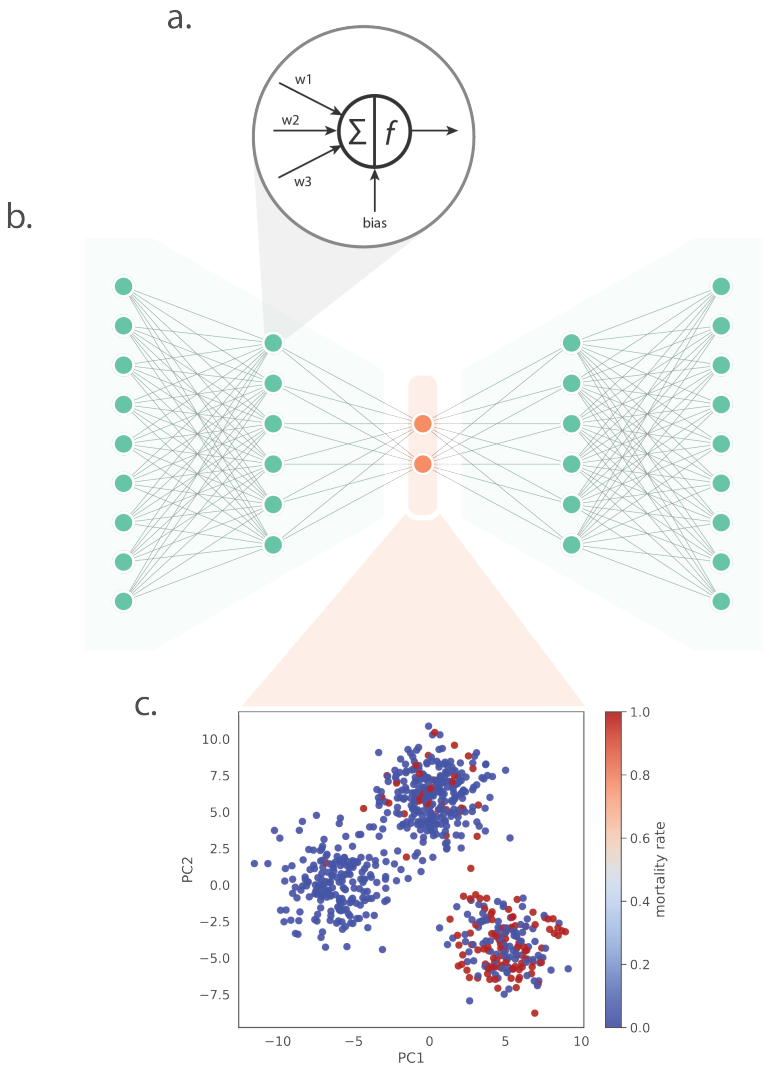


Figure 1.6: **Example of a difficult medical problem**. (a) Computations within a single neuron. Input signals from successor nodes with their corresponding weights (*w1, w2, w3*) as well as the bias are received. Subsequently, an activation function (*f*) is applied to their weighted sum ($\sum$). (b) Structure of an autoencoder (AE). The encoder (left) consists of contiguous hidden layers, each with fewer nodes. The latent embedding (yellow) represents the bottleneck of the AE with the minimum number of nodes. The decoder reversely mirrors the layer structure of the encoder, with the final layer featuring the same number of nodes as the input layer as it attempts to reconstruct the original input from the latent embedding. (c) Exemplary clustering of GBM patients, coloured by mortality rates. Three clusters are visible, with one showing visibly higher overall mortality rates. Graph generated using simulated data.

therefore utilise a special type of artificial neural network: the autoencoder.

Autoencoder (AE) are neural networks designed for unsupervised learning (Figure 1.6b). Autoencoders show a mirrored architecture, consisting of an encoder and a decoder part. While the encoder maps the high dimensional input data into a lower dimensional latent embedding, the decoder attempts to reconstruct the original input from said embedding. The network is trained to try to minimise the error between original input and reconstruction [65]. The unique architecture of autoencoders opens up intriguing possibilities for various applications. For instance, their encoder-decoder structure allows autoencoders to act as efficient dimensionality reduction tools capable of accommodating input data in a joint low-dimensional embedding. As opposed to other dimensionality reduction methods like principal component analysis (PCA), autoencoders are able to model non-linear feature interactions and accommodate heterogeneous inputs from different sources, making them a valuable tool for data integration in precision medicine [84]. The low-dimensional latent embedding can also be adapted for generative purposes in various ways, for example by incorporating probabilistic elements into the encoding process. This variant, known as the *variational autoencoder (VAE)*, enables sampling from the latent space, allowing the generation of diverse outputs for a given input [85]. A semi-supervised extension of autoencoders termed *conditional variational autoencoder (cVAE)* have recently gained popularity in genomics due to their potential to guide the learning and generation of latent representations. By supplying auxiliary information to the encoder and decoder (e.g. age of patients) the model's behaviour can be controlled. This "conditioning" of autoencoders has shown promise as an effective tool for eliminating undesirable signals from biomedical data, such as measurement errors or technical noise [86, 87].

Overall, deep learning models possess remarkable versatility, capable of handling diverse datasets by avoiding assumptions about data structure and leveraging the ability to approximate complex interactions through layer stacking and neuron combination. However, despite their immense power, the application of deep learning models encounters significant challenges, notably limiting their utility in medical use cases [88].
To illustrate this, let us once again examine our clustering of GBM patients. After successfully training an autoencoder on the measured DNA methylation data, we retrieve three distinct cluster of patients, indicating different clinical sub-types. Mapping the mortality rates of patients belonging to each cluster reveals that one cluster shows a significantly increased mortality (Figure 1.6c). From a clinical point of view, it is of paramount importance to now identify DNA methylation biomarkers in order for these patients to be stratified early to improve their outcomes.
This means, we have to start understanding according to which criteria the autoencoder

groups together patients, to learn about biomarkers characterising patients belonging to each cluster. In other words, we want to explain or even fully interpret the latent space. This seemingly simple task, however, is not trivial. As the architecture of deep learning models becomes more complex, with numerous layers and parameters, they become inherently more difficult to interpret, ultimately resembling ”black boxes” [88]. The interpretation of how and why deep learning models make specific predictions, or cluster specific patients together, thus poses a significant challenge. So we conclude that to bring AI models out of research and into a clinical setting, we need to focus on increasing the transparency of models.

1.2.3 Towards Trustworthy AI

> *The First Law*: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
> *The Second Law*: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
> *The Third Law*: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
> *The Zeroth Law*: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
>
> *Isaac Asimov, I, Robot, 1950*

As AI-powered applications continue to impact our daily lives due to their use in high-stakes applications like healthcare, business, government, education, and justice, the societal awareness regarding the safety, reliability, and overall trustworthiness of these applications has escalated [89, 90]. While researchers and developers commonly prioritise enhancing the predictive performance of their models, solely focusing on accuracy is no longer adequate. Rather, new requirements arise, prioritising the reliability of models over their accuracy. Factors contributing to a model's trustworthiness include, but are not limited to: explainability, interpretability, transparency, fairness, robustness, stability, responsibility, accountability, and ethics [91]. Although ideally, a trustworthy model should encompass all the aforementioned factors, in this work, I will concentrate on a subset of the three

most central and crucial aspects: explainability, interpretability, and fairness.
There exists a lack of consistency between studies in using the terms "interpretable" and "explainable" AI [91]. However, it is commonly acknowledged that there does exist a conceptual difference between the two which should not be confused [90]. Therefore, the following notations will be used in this thesis for consistency and clarity:

- **Explainability**: ability to understand **why** a model derived a final prediction. Explaining a model is the most important aspect when communicating results to end-users, i.e. clinical personnel, as it gives insight into the considerations (and as a logical conclusion also potential pitfalls) of the model.
- **Interpretability**: ability to fully understand **how** the model functions. This implies understanding the "inner workings" of a model, which is also referred to as intrinsic interpretability. By this definition, interpretability captures more detail than explainability, as it aims for full transparency of the model.
- **Fairness**: attempt to correct for any algorithmic bias in the decision processes of models. Especially models used for societal purposes commonly include "sensitive" variables, including information on gender, race, disabilities, and age [92]. Recent years have seen an increased awareness of the pitfalls and limitations of these models, with disturbing reports on tools used for recruiting [93] or juristic purposes [94] exhibited discriminatory behaviour against subgroups, including women and darker-skinned individuals [92, 95]. A recent large scale study on ICU data reported disparate treatment in prescribing mechanical ventilation among patient groups across ethnicity, gender and age, proving that also medical applications are not free of unwanted bias [96]. This unfairness of models is rooted in the data-driven learning of algorithms, as training data might contain human biases, which must be actively corrected for [95].

Why Strive Toward Explainable, Interpretable, and Fair Models?

So, why exactly we should aim at making models ideally interpretable, but at least explainable, and always fair?

Clinicians' needs As already illustrated by the practical scenarios (section 1.2.2) the limited interpretability of the decision-making of AI models stands in the way of successful adoption in clinical setting [97–99]. A recent survey attempting to capture the expectations clinicians have with regard to machine learning explainability revealed that they "view explainability as a means of justifying their clinical decision-making" [100]. Interestingly,

and in opposition to the technical researchers in the field who strive for evermore accurate predictions potentially at the cost of explainability, the survey revealed that for clinicians it is not all about accuracy - even models falling short in performance were deemed acceptable as long as they provided clarity on why the model under-performs in this specific case. This also means that the model does not necessarily be intrinsically interpretable, but rather it should permit a decision of trust, rather than trust itself [101].

Scientific Validation and Discovery

> Ideally one might marry people's unique knowledge of what is comprehensible with an algorithm's superior capacity to find meaningful correlations in data; to have the algorithm *discover* new signal and then have humans *name* that discovery
>
> *J. Ludwig and S. Mullainathan, 2023* [102]

Upon the breakthrough of deep learning in multiple fields, some (overly optimistic) researchers believed the end of classical, hypothesis-driven science to be near. In the future, the believe, all novel insights would come from the unsupervised algorithmic analysis of large datasets [15]. As we know today, this was not - and will probably never be - the case, as the correlation patterns AI algorithms uncover in data do not necessarily imply causation. Nevertheless, it is worthwhile for researchers to seek algorithmic explanations in the form of interpretable models for two reasons: validation and discovery [103].

1. **Validation of performances:** Firstly, explainability can be considered a way of improving a model's accuracy, as it highlights the limits of the model that can potentially be improved through training data used or additional parameters, ultimately enhancing the predictive power [90]. Secondly, introducing explainability into models is an effective guard against unexpected errors and can improve generalisability by preventing overfitting. For instance, let us consider a model developed for the stratification of patients in high-risk and low-risk groups in an ICU setting [104]. Counter-intuitively, this model classified patients with clearly increased fatality risk, due to i.e. comorbidities, as having a low risk of mortality. Subsequent interpretability analyses revealed that this false prediction derived from a spurious association in the training data of the model: respective patients were at such great risk that they received extra medical attention, effectively lowering their risk of mortality.

One can only imagine the grave consequences of using such a model without expert supervision in the decision-making in a triage setting of a hospital.

2. **Discovery of underlying biological mechanisms:** To illustrate how interpretable models can be used to expand our biological knowledge, let us assume we were successful in developing a neural network which shows high performance in the diagnosis of a rare disease. Due to the rarity of the disease, paired with its unusual molecular mechanisms, we have limited knowledge of the disease's molecular pathophysiology. If our neural network proves to be not only performant but also robust and replicable, this suggests that it bases its decision-making on true, previously unknown signals. This true signal might not have been captured by simpler methods (linear regression) as it is e.g. the non-linear combination of multiple variables. By incorporating interpretability into our respective model, we can unravel its decision-making process, enabling a deeper understanding of the molecular mechanisms that distinguish healthy from diseased individuals. This not only enables us to gain insights into patient stratification but also empowers us to generate new hypotheses for further investigation. The recently published framework "Geneformer" exemplifies the range of possibilities such a framework offers for hypothesis generation: through *in silico* perturbations they explore the downstream effect of gene mutations or propose novel candidate therapeutic targets [105].

"**The Right to Explain**" It is however not only the opinion of clinicians that requires researchers to consider the transparency of their model when developing AI algorithms for health-related settings. Since the late 2010s, lawmakers and international authorities continuously established requirements regarding the reliability of AI models. Early efforts of the European Commission include the establishment of an independent High-Level Expert Group on Artificial Intelligence (AIHLEG) in June 2018, which published an assessment list for evaluating AI systems regarding their trustworthiness [106]. The work of this group has later been expanded in the form of an official European Commission's Artificial Intelligence Act (EU AIA), which is deemed the world's first comprehensive AI law regulating the use of AI in the EU. Specifically, its goal is to "ensure that AI technologies are utilised ethically and responsibly, aligning with fundamental rights and societal values" [107], of which the model's fairness is an integral part. Furthermore, article 13 of the EU AIA explicitly highlights the need for medical applications to be explainable and/or interpretable by stating that "*high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable providers and users to reasonably understand the system's functioning*" [108]. Also, the General Data Protec-

tion Regulation (GDPR) has seen recent additions concerning the use of AI, emphasising "explainability" as a top priority in machine learning research, phrasing the right of data subjects (i.e. patients) to "*obtain an explanation of the decision reached*" and to "*challenge the decision*" in the special case of decisions reached using automated processing [109]. All of these recent changes to international policies make it clear that developing AI models that are transparent, explainable, fair, and overall trustworthy is no longer an option, but a must, which is enforced through EU law.

Glimpsing Inside the Model: From Black to (Partially) White Box

As the architecture of deep learning models becomes more complex they become inherently more difficult to interpret and are therefore commonly described as resembling "black boxes". This terminology has become prevalent in literature, leading to the classification of models as white, grey, or black boxes based on their degree of interpretability.

White box describes models that offer complete interpretability, allowing users to fully understand their inner workings, decision-making processes, and parameters. These models provide clear insights into how inputs are transformed into outputs, making it easy to trace and interpret their behaviour. Mechanistic models used in molecular modelling, rule-based models, and simpler machine learning models such as linear and logistic regression, decision trees, and Random Forests with a limited number of trees are generally considered to be white box models, as they build to be intrinsically interpretable [110, 111].

Grey boxes combine the advantages of inherently interpretable modelling techniques like mechanistic models (white box) with the predictive power of complex machine or deep learning models (black box). For example, neural networks can be included in mechanistic models to account for systematic errors or estimate unobservable parameters [112]. Recent advancements in deep learning for genomics include the utilisation of sparse neural networks, which are black box models with a simplified architecture making them fully or at least partially interpretable (see section1.2.3)

Black box are models that do not offer any insight into their decision-making. These models typically provide accurate predictions or classifications but due to the large number and diversity of their layers, they offer little insight into how those results are generated [113]. Input features important for derived predictions can only be approximated through explainability methods (see section 1.2.3) Densely connected deep neural networks are a common example of a black box model.

It is a common conception that the degree of interpretability of a model is indirectly correlated with its predictive power (Figure 1.7). This means, that the more a model resembles a

white box, the lower its performance and vice versa. This paradigm of a trade-off between interpretability and accuracy in models introduces a layer of complexity for researchers striving to enhance the transparency of their models. But which methods do biomedical researchers have at their disposal for making their models interpretable, or at least explainable? In the following section, I will highlight common methods used in genomics to achieve transparency in models. A more extensive description of interpretable genomics can be found in **Chapter 4**. I will differentiate between *passive* and *active* interpretability methods, where the former is model-independent and can be readily applied after model training, while the latter requires changes to the architecture of the models.

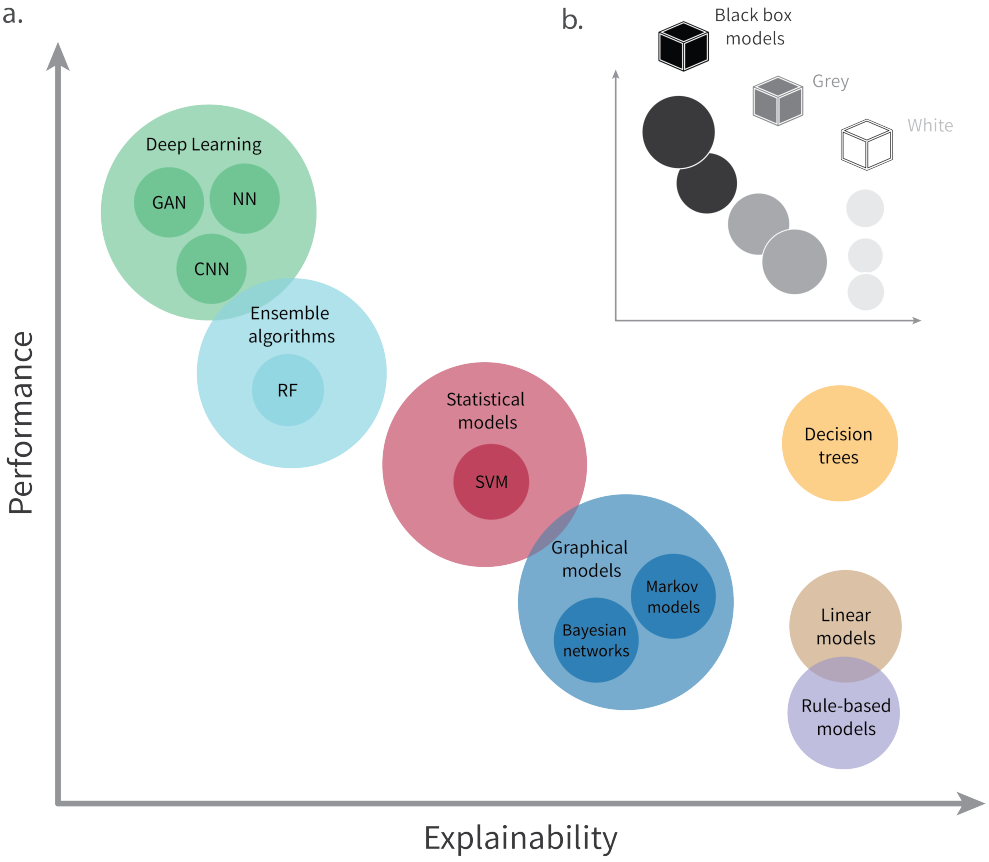


Figure 1.7: **Relationship between model explainability and performance.** (a) Trade-off between explainability and peformance for a selection of machine learning and deep learning models. (b) Intuition on which category of models can be regarded as black, grey, or white boxes. Figure adapted from 'Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond', by Guang Yang *et al.*, in *Information Fusion*, 2022, Elsevier [114]. GAN: Generative Adversarial Network, NN: Neural Networ, CNN: Convolutional Neural Network, RF: Random Forest, SVM: Support Vector Machines

Passive Explainability: Post-Hoc Attribution Approaches Passive attribution approaches stand out as the most popular and extensively employed approaches for enhancing model transparency. Their popularity stems from their inherent model-agnostic nature and the convenience of application after model training (post-hoc). Given the widespread adoption of attribution approaches, a multitude of diverse strategies have emerged, each offering distinct methodological implementations. Here, I will focus on two of the most prevalent strategies that also make an appearance in this thesis: perturbation/permutation and game theory. I will give an intuition on their underlying concepts, as well as highlight advantages and limitations.

Permutation as well as *perturbation* of input features belong to the earliest explainability methods, with Garson's *et al.* sensitivity analysis having been published in 1991 [115]. Permutation methods involve the random shuffling of features to disrupt their connection to samples, while perturbations maintain the data's structure while altering individual feature values. Both methods forward-propagated these changes and analyse the resulting changes in the network, comparing alterations to the original data to determine the significance of features [116, 117] and even hidden nodes [118–120]. While the straightforward concept of perturbations makes them a universally appreciated method, their practical utility is hindered by scalability issues. Specifically, the computational burden associated with generating a large number of perturbations limits their effectiveness, particularly when applied to a large-scale dataset or when investigating combinatorial feature interactions.

SHAP (SHapley Additive exPlanations) [121] is founded on principles from game theory and provides local approximators by looking at the model as a cooperative game, where players are input features (e.g. genes) and credits are the model's prediction. It seeks to fairly allocate credit among the input features responsible for the model's output, assigning more credit to more important features. The derived SHAP values are easy to interpret and utilise, which has contributed to its widespread adoption as one of the most commonly used methods in explainable machine learning. However, the utility of SHAP for large genomic datasets is limited due to its high computational expense. Generally, computing exact SHAP values is considered NP-hard, and although approximation algorithms exist, they do not scale well to the large datasets prevalent in biomedical sciences.

Active Interpretability: Sparse Neural Networks As the name implies, sparse or visible neural networks, introduce interpretability by sparsifying the architecture of networks, thereby significantly reducing their complexity. The approach is gaining momentum and interest in the genomics community, as sparsification does not occur randomly, but is achieved by integrating biological knowledge into the network. In the resulting biologically

informed sparse neural networks, every node represents a gene, pathway, or even tissue, while edges represent the known interactions between them. The reduced number of interactions also makes the model (intrinsically) interpretable and allows researchers to directly deduce relevant genes or pathways from a model decision [122–125].

1.3 Aim and Outline of the Thesis

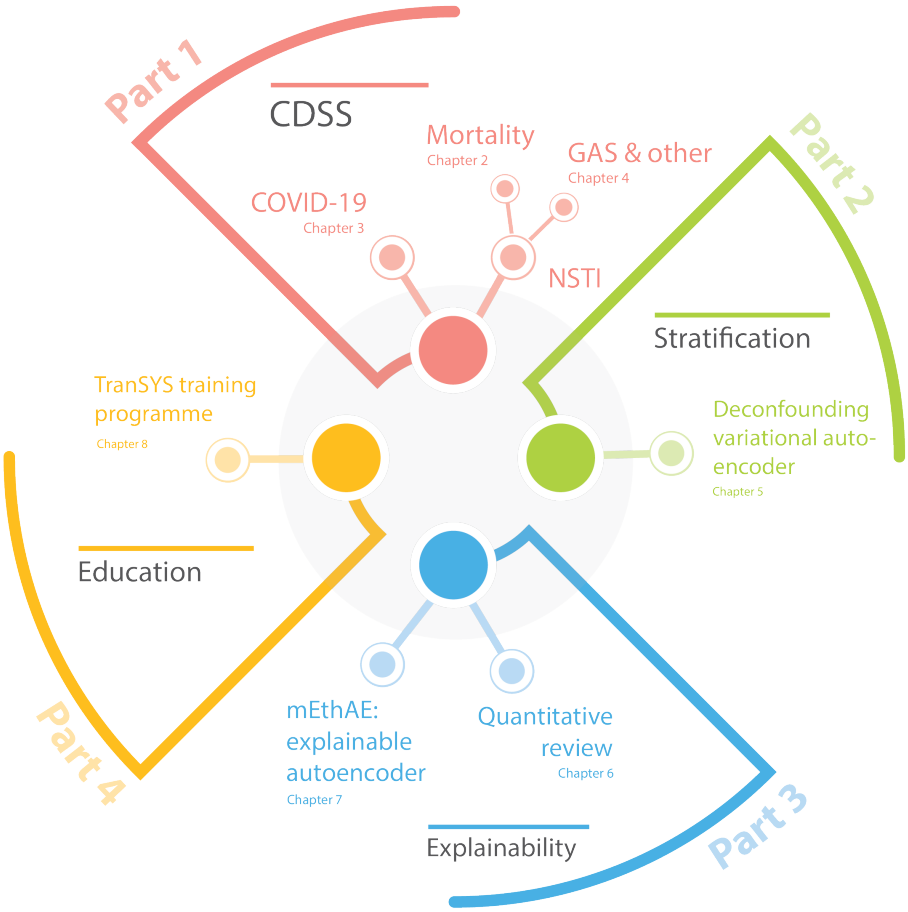


Figure 1.8: **Structure of this thesis.** Each Part of the thesis focuses on a different aspect of precision medicine: Clinical Decision Support Systems (*Part 1*), patient stratification (*Part 2*), model explainability (*Part 3*), and interdisciplinary education in precision medicine (*Part 4*). Omitted are the *Introduction* and general *Discussion*. CDSS: Clinical Decision Support Systems ,GAS: Group A Streptococcus, NSTI: Necrotising Soft Tissue Infections

Throughout this chapter, I have explored the multi-faceted field that is Precision Medicine, offering insights into its diverse computational aspects. Within each facet, I highlighted the associated challenges hindering its seamless integration into clinical practice. These obstacles range from handling the vastness, heterogeneity, and complexity of biomedical data, to navigating through the plethora of analysis algorithms or developing models that are not only accurate but also deemed trustworthy.

This thesis aims to advance the integration of precision medicine principles into healthcare by developing artificial intelligence models utilising a variety of biomedical data and focusing on explainability and fairness. By prioritising transparency in models, I sought to create innovative tools that are not only effective but also capable of providing insights into the underlying biology of diseases while meeting clinicians' requirements.

I achieve this aim by:

(I) developing machine learning-based clinical decision support tools.

(II) designing fair stratification tools for cancer patients that integrates information from diverse biological layers.

(III) utilising explainable deep learning algorithms to explore the complexities of the human epigenome.

The **Introduction** offers an overview of the concept and objectives of precision medicine, providing concrete medical examples to illustrate how our healthcare system can benefit from this novel paradigm. It emphasises the central role AI plays in enabling precision medicine and introduces the algorithms that will be discussed in the thesis through hypothetical yet realistic case studies. Furthermore, it underscores the importance of understanding the decision-making processes of AI models and describes approaches that enable the explanation of these models.
In **Part 1** I explore how we can utilise precision medicine concepts to enhance the management and care of patients suffering from Necrotising Soft Tissue Infections (NSTI) (**Chapter 2 and 4**) and COVID-19 (**Chapter 3**). Working in close collaboration with clinical experts, I developed machine learning-based tools to predict important patient outcomes, such as the risk of mortality, or the anticipated disease severity. Besides attempting to accurately predict these outcomes, I focus on the explainability of models, closely analysing which medical factors determine their decision-making which ultimately enables clinicians to make more informed decisions.
Part 2 delves into the world of disease sub-typing and patient stratification. While grouping patients using high-dimensional, multi-omics data is a widely adopted approach, the

impact of confounders - external factors unrelated to the condition, such as batch effects, age, or sex - on clustering is often overlooked. In **Chapter 5**, we thus propose and compare deep learning methods for achieving fair patient clustering, which involves clustering that is devoid of confounding factors.

The focus of **Part 3** lies in demonstrating how to use explainable deep learning algorithms for hypothesis generation in genomics. The recent years have seen the development of numerous diverse interpretability strategies, making it increasingly difficult to navigate the field. Therefore, the first part of the chapter, **Chapter 6**, provides an overview of the current state of the field. By employing predefined criteria, we identify the interpretability solutions most commonly utilised, highlight exceptional examples, and pinpoint promising areas for further research. **Chapter 7** subsequently illustrates the use of explainable AI in exploring novel biological associations. By combining explainable deep learning with epigenetic data, we illustrate the potential of not only efficiently modelling the input data but also uncovering interactions lacking extensive characterisation.

The final part of this thesis, **Part 4**, explores the interdisciplinary nature of precision medicine. Addressing the challenges associated with the paradigm shift in disease prevention and healthcare necessitates the integration of expertise from various fields, including genomics, informatics, medicine, and ethics. Training the next generation of precision medicine experts thus requires the development of cross-disciplinary, international, and education-focused doctoral programs. In **Chapter 8** we examine the training and collaboration experiences of early-stage researchers within TranSYS, an Innovative Training Network (ITN) for precision medicine funded under the EU's Horizon Europe program. To strengthen these education-focused doctoral programs, we identify challenges and propose solutions to design improved training networks for the next generation of aspiring researchers.

The thesis ends with a **Discussion**, which summarises key findings and places them within the broader context of precision medicine. A significant portion of the Discussion is devoted to highlighting the open challenges that need to be addressed to integrate precision medicine into routine clinical practice. Lastly, I provide my perspective on opportunities that are essential in shaping the future of the field. The organisation of this thesis is visualised in Figure 1.8.

1.4 Supplementary Notes

Acknowledgements

Language, expression, and grammar in this chapter were polished using ChatGPT to improve readability.

Part 1

Clinical Decision Support Systems

Chapter 2

Decision support system and outcome prediction in a cohort of patients with necrotizing soft-tissue infections

Sonja Katz[†], Jaco Suijker[†], Christopher Hardt, Martin Bruun Madsen, Annebeth Meij-de Vries, Anouk Pijpe, Steinar Skrede, Ole Hyldegaard, Erik Solligard, Anna Norrby-Teglund, Edoardo Saccenti, and Vitor A.P. Martins dos Santos

[†]authors contributed equally

Published in: *International Journal of Medical Informatics. 2022 Nov 1;167:104878*
DOI: 10.1016/j.ijmedinf.2022.104878

Abstract

Background: Necrotizing soft-tissue infections (NSTI) are severe infections with high mortality caused by various pathogens affecting a heterogeneous population. Thus, there is the need for a Clinical Decision Support System (CDSS) able to detect NSTI early, to provide an overview of expected outcomes and individualized treatment recommendations, thereby also increasing early awareness among clinicians.

Methods: Interviews with eight clinicians from different departments were performed to identify relevant clinical needs, resulting in 24 unique questions. Interesting clinical questions were subsequently used as targets to develop a machine learning based predictive system (Random Forest). Data obtained from the INFECT study, comprising 409 prospectively included NSTI patients [126], was used.

Results: Risk of 30-day mortality was generally deemed relevant in interviews and thus selected as the primary health endpoint. It could accurately be estimated using sixteen parameters available in the data set on the first 24 h following ICU admission. These parameters mostly reflected the severity of sepsis (i.e. lactate, urine production, blood pressure) and one baseline characteristic (age), while no NSTI specific parameters (location or type of NSTI, skin changes) were found predictive. The ROC AUC of 0.91 (95% CI, 0.88-0.96), displayed better performance than the SOFA score (ROC AUC=0.77 (95% CI, 0.69-0.84), p-value=$5.07E{-}05$) and was slightly higher than SAPS II prognostic estimates (ROC AUC=0.88 (95% CI, 0.83-0.92)). The models proved to be robust regarding missing input parameters, with ROC AUC > 0.8 even in the case of $>$50% missingness. Subsetting of selected variables according to typical clinical availability revealed the possibility to make reliable predictions (ROC AUC > 0.8) using data obtainable within the first hour (vital parameters, blood gas values) in the ICU.

Conclusions: This study lays the foundation for a CDSS in NSTI. It indicates that risk of mortality can be accurately estimated for NSTI patients utilizing a machine learning based approach. By extending predictions to other relevant characteristics (i.e. risk of septic shock, risk of acute kidney injury) a complete, clinically relevant overview of the expected disease course can be obtained. Ultimately, this can support clinicians in making early treatment choices, resulting in improved resource management and clinical outcomes.

2.1 Background

Necrotizing soft-tissue infections (NSTI) are rare, fulminant infections, which affect heterogeneous population in regards to age, sex, and the presence of comorbidities [127]. Both a single pathogen (monomicrobial type) or multiple pathogens acting synergistically (polymicrobial type) may be responsible, with different pathogenic mechanisms [128]. Besides local tissue destruction, these pathogens cause systemic toxicity, leading to sepsis, and in many cases septic shock (28-50%) [129, 130]. If left untreated, NSTI will be fatal within days, making timely recognition an essential prerequisite for successful disease management. Current reported mortality is 10-29% [129, 131–135]. Besides mortality, long-term morbidity is extensive: functionally due to scars, amputations, fatigue, as well as psychosocially, which may including fear for recurrence, post-traumatic stress, depression and changes in social activities [136–139].

2.1.1 Treatment

As soon as the diagnosis of NSTI is suspected, surgical inspection is initiated, followed by obtaining tissue cultures and surgical debridement when the diagnosis is confirmed. Broad-spectrum intravenous antibiotic therapy should be started, as well as resuscitation and organ support in the ICU in most cases (68-92%), according to the severity of the sepsis [132, 140, 141]. Amputation of extremity needs to be performed, usually if muscular involvement is so extensive that no functionality of the affected body part is expected [142]. Both intravenous immunoglobulins (IVIG) and hyperbaric oxygen treatment (HBOT) may be used additionally.

A major challenge in the effective treatment of NSTI is its early recognition due to the low incidence of NSTI, as well as the often non-specific symptoms upon presentation. Reportedly, 41-96% of patients with NSTI are misdiagnosed upon presentation [143]. After the diagnosis is established, another major challenge is to provide adequate, individualized treatment, as early as possible, while preventing over-treatment or under-treatment. Current treatment strategies are empiric at the start, with refinement taking place later as the disease progresses. For example, initially broad-spectrum antibiotics are administered until the definitive cultures are known, which can take multiple days. IVIG administration usually depends on the presence of septic shock in combination with gram stain results, which may take several hours from presentation to become known [129]. Early predictions of streptococcal involvement and risk of septic shock could therefore lead to earlier introduction of targeted treatment. This might also apply to mortality, where differentiating those with a high risk of death versus those with high chances of survival, could help to

allocate resources earlier and more efficiently, thereby potentially reducing both over- and undertreatment.

2.1.2 Predictive scores

Currently, there are various prediction scores for NSTI or sepsis in general. Diagnostically, the LRINEC score was developed in 2004, to identify patients having a high risk of NSTI based on laboratory values [144]. However, despite promising initial results, a recent meta-analysis of observational studies using this scoring system demonstrated low sensitivity for diagnosis of NSTI [145]. The prediction of outcomes, which is important for both treatment allocation as well as in the communication with patients and their families, is currently performed for mortality using general ICU mortality predictions (SOFA score [146], SAPS II [147], SAPS III [148], APACHE IV score [149]). Also other prediction systems are available, for example for estimating Acute Kidney Injury (AKI) [150, 151] or the risk of septic shock development [152]. It has however not been sufficiently studied how well these general prediction scores perform among patients with NSTI. In a mixed ICU population for example, one study found the SAPS II mortality prediction to be less in case of sepsis compared to other disorders for which patients were admitted [153]. Also, the requirement of different predictive scoring systems to estimate different outcomes is not practical. Ideally, a more comprehensive overview of expected disease characteristics, disease progression, and outcomes would be obtained early after admission. For example, the ability to accurately estimate whether streptococcal involvement is likely and whether sepsis may progress to septic shock, could result in earlier administration of IVIG, to ideally reduce progression to severe sepsis. Although microarray based analysis systems promise to reduce the time from sampling to bacterial identification to only few hours, such systems are still under development and thus not yet readily available in clinics [154].

2.1.3 Clinical decision support system

We believe the outlined shortcomings in efficacy (accuracy, applicability) of general clinical scoring systems might be addressed by a so-called clinical decision support system (CDSS). A CDSS is a framework aiming to link health observations (e.g. clinical data, patient-reported outcomes) with health knowledge, thereby supporting the decision-making process by providing consultation to medical personnel. It consists of multiple levels, incorporating data management, modeling, and data visualization in one platform. The central piece of the CDSS, the predictive models, are frequently based on machine learning or deep learning methods and utilize Big Data approaches to identify disease, disease stages,

as well as patient-specific patterns. They have shown promising results in the diagnosis of sepsis [155, 156], and gain popularity as personalised medicine tools [157]. A CDSS holds many advantages compared to the current scoring systems in use. They can be designed to be highly versatile, addressing different health endpoints at the same time, such as treatment choices or prognostic outcomes. Creation of such a multiple-endpoint CDSS abrogates the need for individual scoring systems used in parallel. Furthermore, CDSS are flexible regarding the use of available information and can be tailored towards one specific disease. Another advantage is their circular, iterative design. By building the predictive models on top of a database system, new patients can easily be added or available information updated in an automated manner. This, in turn, makes predictions more and more accurate - so the system learns as it grows. Decision support in the treatment of NSTI has until now been limited to simple algorithmic procedures in guidelines [158], with substantial differences. Some of these differences include the timing to second look (<12h [159], 24-48h [160], 48-72h [161]), use of HBOT (never [161], consider if available [160, 161]) and IVIG suppletion (not mentioned [162], in case of GAS [159], in case of organ dysfunction [160]). A CDSS is expected to improve evidence based treatment, both by being used complementary to local guidelines, or by use on its own.

In this publication, we present the first steps towards the realization of a multiple-endpoint CDSS for improving NSTI patient care. To identify the most relevant clinical shortcomings in NSTI care and outcomes, we first conducted interviews with clinical specialists involved in the diagnosis and treatment of NSTI. As proof of concept and to illustrate how a CDSS could assist the decision-making process, the risk of mortality for patients was predicted. In an attempt to compare our CDSS to current systems in use, we benchmarked its performance on the SAPS II and SOFA score at admission.

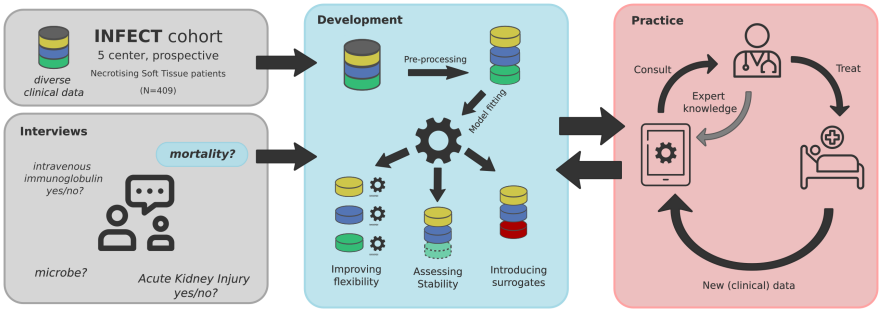


Figure 2.1: **Graphical abstract.**

2.2 Methods

2.2.1 Semi-structured interviews

To create an overview of the various relevant clinical questions, semi-structured interviews were conducted (Figure 2.2, Table S2). An interview guide was constructed with the PerMIT project group, an international consortium dedicated to demonstrating the potential benefits of personalized medicine approaches for NSTI and sepsis patients (https://permedinfect.com/, grant number 8113-00009B). Eight clinicians (3 ICU specialists, 2 surgeons, 1 microbiologist, 1 ER specialist, 1 general practitioner) were interviewed by one of the authors (JS). During the interviews, participants were asked which clinical questions they believed were most relevant in the various phases (pre-hospital, pre-ICU, ICU). When new questions emerged, those were added. Each participant was asked to attribute a score for relevance to each question; (3) highly relevant, (2) relevant, or (1) interesting but not relevant. In case of insufficient knowledge on a topic, it could be left blank. Some questions were added later in the process, in which case fewer participants scored it for relevance. The average score for the relevance of each of the questions was calculated for those that attributed a score to a question.

2.2.2 INFECT study cohort

The INFECT data set utilized in this study is the result of an international, multicentre, prospective, cohort study of adult patients with NSTI included prospectively at five Scandinavian hospitals, which were referral centers for NSTI (INFECT study: ClinicalTrials.gov, number NCT01790698). A total of 409 patients above the age of 18 and with surgically confirmed NSTI cases were enrolled [126]. Data recorded in the INFECT study include patient demographics, clinical data (blood samples, clinical findings), daily ICU data for a period of up to 7 days (fluid administration, medication, observed parameters), information regarding specific treatments and samples (surgical procedures, microbiological findings, HBO treatments), as well as follow-up data (90-day follow-up, 365-day follow-up). All of this information is encoded in a total of 2,400 variables, making the INFECT study the largest prospective study in patients with NSTI to date. A detailed description on the INFECT cohort is offered by Madsen et al. [126].

2.2.3 Data pre-processing

Time-dissection of data set

To identify the earliest possible time-point for mortality predictions, the data set was split into 11 subsets. Respective subsets comprised: ENTRY (upon hospital admission), PRE-SURGERY (prior to first surgical procedure), POST-SURGERY (posterior to first surgical procedure and prior to ICU admission), BASELINE (BL; first 24 h of ICU admission), ICU-day1 (first day in ICU), ICU-day2 (second day in ICU), ICU-day3, ICU-day4, ICU-day5, ICU-day6, ICU-day7. The ICU day follows the fluid charts, typically from 06.00 to 06.00. Day 1 is from admission time to the start of the fluid chart the next day, giving variable lengths of day 1 from 0 to 24 hours. To address this discrepancy between patients, the BASELINE data set was included, as variables therein pertain to time of ICU admittance and 24 hours forth.

Data cleaning & imputation

The selected (binary) target was 30-day mortality. All other clinical endpoints were omitted, as were patients with missing mortality data (n=4), resulting in a sample of 349 alive and 56 deceased patients. Only data available upon ICU admission were included. Irrelevant or potentially biasing variables (e.g. PatientIDs, dates), duplicates, or variables with disputable accuracy due to subjectivity of the data (i.e. estimated Glasgow Coma Scale) were additionally removed from the analysis. Imputation was performed if the total number of missing entries per variable did not exceed 5%, while variables with higher percentages of missing data were discarded. Discarded variables included mostly cytokine measurements, few preoperative lab values (hemoglobin, glucose, lactate, natrium), skin anesthesia and crepitus upon presentation, gas upon radiology, alcohol, and smoking status. To account for the mixed data types present in the data sets, we differentiated between continuous (numerical), binary, and categorical features during imputation. Continuous features were treated using the *IterativeImputer* from scikit-learn [163] and scaled through min-max normalisation. Missing binary information was completed using k-Nearest Neighbours method. Categorical features, such as hospital names, were imputed by the most frequently occurring value and encoded to numerical representation by an ordinal encoder. For the total numbers of variables and patients included in each subset after data cleaning and imputation see Supplementary Material, Table S1.

Variable selection

Relevant variables were selected using a combination of unsupervised filtering and manual curation in a two-step process. Firstly, the preprocessed, time-dissected data sets underwent an unsupervised feature selection step, in which the full feature space was filtered using the python implementation of the Boruta algorithm [164]. Secondly, during optimization of the best-performing model, the filtered variables were manually curated, removing variables that were deemed impractical to use in clinical practice. The original variable names in the INFECT data set, abbreviations used in this publication, and a more detailed clinical description of curated variables can be found in Supplementary Material, Table S2.

2.2.4 Classification

Model development and validation

Random Forest Classifiers (RFC) were utilized to predict patient mortality [74]. Robust internal validation was achieved by an iterative 5-fold double cross-validation (DCV) approach (100 iterations). To quantify the quality of the classification models, we calculated and compared several different metrics, including areas under the receiver operating characteristic curve (ROC AUC), the F_1 score, and the F_2 score. Conversion of SAPS II and SOFA scores to probabilities of mortality was done using the relationship established by Le Gall et al. [165] and Moreno et al. [166], respectively. Confidence intervals of ROC curves were computed using bootstrapping. For the p-values, a two-sided test for difference in AUC was performed. For additional information on model development, validation and metrics please refer to Supplementary Material, Supplementary Methods.

Assessing model stability

To assess the robustness of our systems towards missing variables, iterative random removal of variables was conducted. Therefore, a pre-specified number of variables (between 1 and $m-1$ where m is the total number of variables in the subset) were removed from the data set and the reduced model was trained and validated using the same approach as during model development.

Selection of surrogate variables

Surrogate variables were determined by calculating the absolute Pearson correlation for selected (primary) variables with the whole feature space, excluding cross-correlation

amongst primary variables themselves. The effects on model performance were assessed by replacing missing variables through surrogates, random variables, or removing them from the data set for all patients. Random variables were defined as features with absolute correlations of less than 5% with any primary variable (correlation < 0.05). Model training and validation were carried out using the iterative DCV described above (100 iterations).

2.2.5 Software

For all classification algorithms the implementations available in the scikit-learn Python library (version 0.24.1) [163] were used. ROC curve bootstrapping and p-value calculation was done using the R package *pROC* [167].

2.3 Results

2.3.1 Interviews

The interviews yielded a total of 24 unique questions that were deemed relevant in the diagnostic and treatment process of patients with NSTI. All emerged during the first five interviews. Of these questions, 14 were treatment support questions, and 10 were predictions (Figure 2.2). Most questions (14) concerned the ICU phase, but relevant questions were also identified in the pre-ICU phase (7) and the pre-hospital phase (4). One question, regarding expected microbial etiology, was deemed relevant in both the ICU and pre-ICU phase, and therefore mentioned in both phases with a different score for relevance (Figure 2.2). As can be observed in the table (Supplementary Material, Table S2) and figure (Figure 2.2), there was substantial variation (1.5 - 3.0) in the average scores for relevance attributed to the different questions. Although all questions are relevant to varying degrees, those that can potentially be answered by the available INFECT data set are of most relevance for the initial development of a CDSS on NSTI. Among the most relevant (score >2) of these questions were the prediction of causative microbes, chance of developing septic shock, chance of mortality, and chance of developing Acute Kidney Injury (AKI). Of these relevant endpoints, the prediction of mortality was selected as the first to develop an artificial intelligence based approach within the CDSS.

2.3.2 Earliest time point for prediction of mortality

Comparison of prediction performances showed distinct differences between different time-dissected data (Figure 2.3). Prediction with ICU data sets (BASELINE (BL) - ICU-

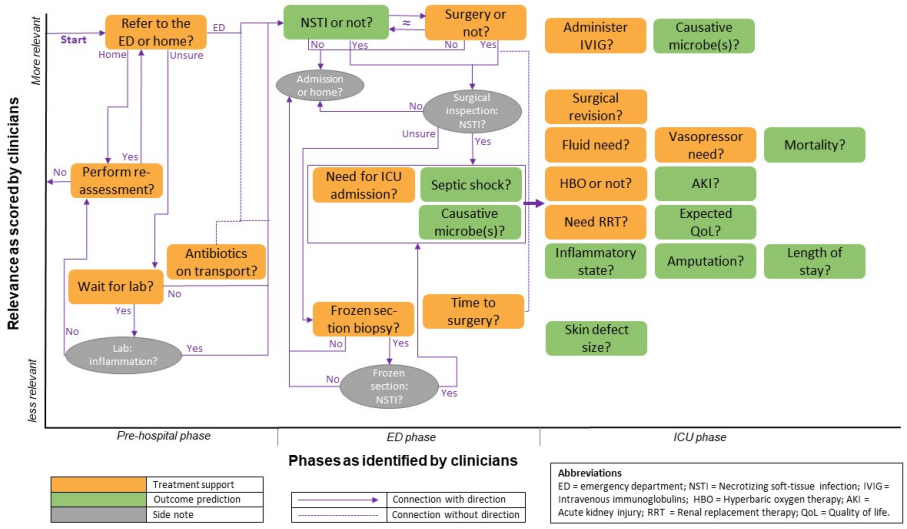


Figure 2.2: **A graphical display of the various (24 unique) treatment support questions (orange) and outcomes (green) which resulted from interviews with 8 clinicians from different involved specialties**. The questions and outcomes are abbreviated, the full questions can be found in Supplementary Note 9. The questions and outcomes are distributed according to phase of the diagnostic and treatment process on the x-axis, and clinical relevance as attributed by the interviewed clinicians on the y-axis. In the first two phases (Pre-hospital phase and ED phase) arrows are placed which indicate how various questions are connected. Grey boxes are added where needed to improve the interdependence of various questions.

day7) revealed that performance peaks using data acquired within the first 24 hours in the ICU (BL), constituting the earliest time point for satisfactory mortality predictions. Therefore, the BL data set was selected to act as the base for further analysis.

2.3.3 Model optimisation

The BL model included a total of 20 variables derived through unsupervised feature selection. To further refine our model, we conducted manual curation, leading to the removal of the variables 'total blood product administration' (impracticality), 'total fluid administration', 'systolic blood pressure' (duplicate entries), 'Glasgow Coma Score (GCS)' (disputable accuracy due to subjectivity of the data), after which 16 variables remained (Figure 2.4A, Figure 2.4B, Supplementary Material, Table S2). Comparison of the predictive power of this model to the SAPS II and SOFA score (Figure 2.4C) revealed excellent discriminatory power (AUC=0.91 (95% CI, 0.88-0.96)) of our system, outperforming the SOFA score (AUC=0.77 (95% CI, 0.69-0.84)), p-value=$5.07E-05$) and showing slightly, yet not statistically significantly, higher AUC than the SAPS II score (AUC=0.88 (95% CI, 0.83-0.92),

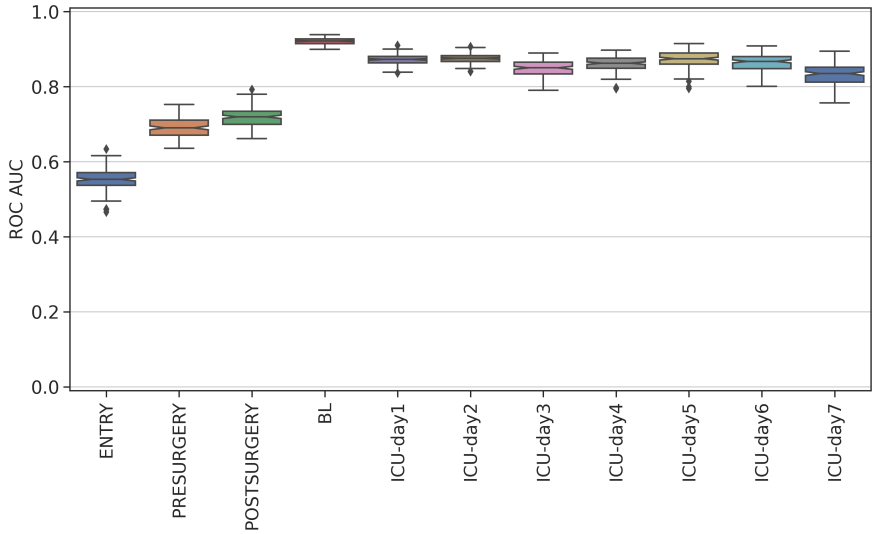


Figure 2.3: **Comparison prediction performances of time-dissected data sets.** Scores displayed as average F1-score (*macro* averaging) over 100 double cross-validation iterations. Notches represent 95% confidence interval around the median.

p-value=0.07).

Probabilities of death were derived and examined for all patients (Figure 2.5A). The calculated likelihood of mortality for patients that are known to be alive (blue) ranges from 0% to around 40% with a dominating peak at around 10%, illustrating the ability of our system to proficiently identify non-critically ill patients. In the probability distributions of patients known to have deceased (orange), the picture is less distinct. Three subgroups of patients can be distinguished, with peaks around 20, 60, and 80% respectively. The lowest-performing subgroup constituted more often of individuals dying at later time points (after day 10), indicating that predictions are better for early deaths (Figure 2.5B). It is evident from Figure 2.5A that using an intuitive threshold of 50% for identifying patients with higher probabilities of death is not optimal and results in elevated numbers of false-negative predictions (Table 2.1). Determination of the optimal threshold (through trying to minimize the number of false negatives) instead yielded an ideal cutoff value of 30% (26%, red line), resulting in a good trade-off between the number of false-positive (FP) and false-negative (FN) predictions.

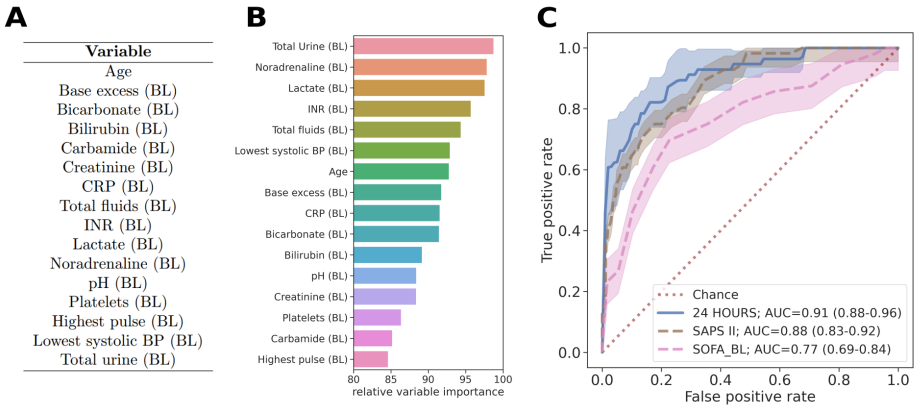


Figure 2.4: **Overview (A) manually curated variables (ordered alphabetically), (B) respective feature importance, and (C) prediction performances of manually curated variables displayed as ROC curve with SAPS/SOFA score as comparison.** The relative importance displayed is the number of iterations (100) minus the (geometric) mean ranking of variables according to their importance during mortality predictions; rank 1 indicated the most important variable and rank 16 the least important.

2.3.4 Improving flexibility and usability

To maximize the flexibility of the developed system the 16 selected variables were grouped into subsets according to their availability (Figure 2.6A). Performance comparison showed that not all identified predictors needed to be included for the model to perform well (Figure 2.6B). Using only variables obtainable within an hour in the ICU (BLOOD set) resulted in a satisfactory mortality prediction. Comparison to the SAPS II and SOFA score proved good discriminatory power with AUC < 0.8 (Supplementary Material, Fig. S1).

In an everyday clinical setting, some variables may be missing. To simulate the effect of missing data a specified number of random variables from our data sets were iteratively removed and subsequently the predictive power of the perturbed set was measured (Figure 2.7). This highlights the stability of our system towards missing information, with ROC AUC scores remaining high (AUC 0.9) even when removing more than half of the data (e.g. removing variables: AUC = 0.89 (95% CI, 0.88-0.89)). Similar results were obtained when using the early available BLOOD set (Supplementary Material, Fig. S2). The gradual increase of standard deviation can be accredited to differences in the importance of variables (Figure 2.4B). With higher missingness rates, important features in the successful

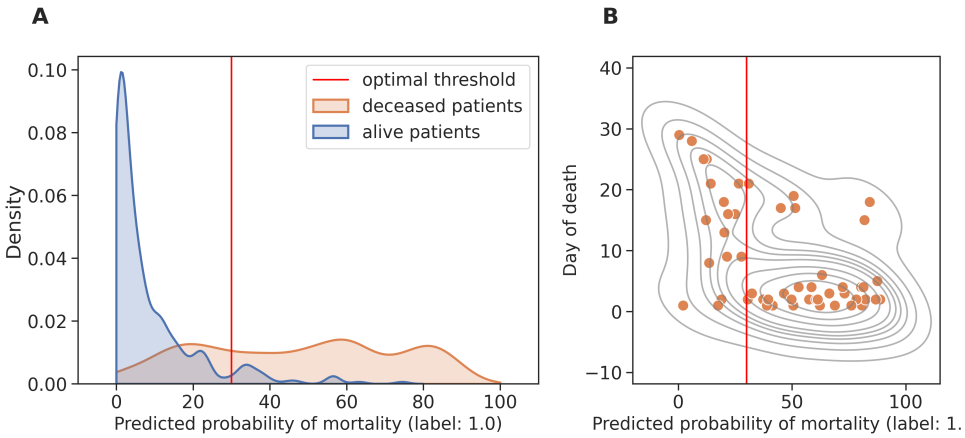


Figure 2.5: **Calculated probabilities of patient mortality.** (A) CDSS predictions for patients known to be alive (blue) or deceased (orange). (B) Patients known to be deceased, plotted with regard to their time-point of death. Red line indicates the optimal decision threshold when aiming to minimise false-negative predictions. Densities calculated from the average probability of mortality for each patient over 100 iterations.

prediction of patient mortality are more likely to be absent, and subsequently prediction performance decreases. Despite the robustness of our system, excluding (key) variables like urine, lactate, or amount of administered noradrenaline, has a negative impact on model performance. Therefore, alternative measurements replacing potentially missing variables were identified (Table 2.2). Exchanging variables with their highly correlated alternatives revealed no noticeable loss of performance, illustrating the relevance of the approach (Supplementary Material, Fig. S3). In the case of multiple missing variables, compensation for absent values performs better than simply removing the missing values from the model. This is of special importance in scenarios of high missingness rates (> 60%) (Supplementary Material, Fig. S4).

2.4 Discussion

This study identified multiple relevant clinical questions and predictions for which data-driven support may facilitate early individualized treatment for NSTI patients. One of the most relevant outcomes to be predicted, identified through interviews, was risk of 30-day mortality, which was subsequently taken as health endpoint to develop a machine learning based predictive system. Mortality predictions proved to be highly accurate, robust, and able to handle substantial amounts of missing data.

Table 2.1: **Summary of model accuracy at different threshold levels.** Highlighted in bold is the threshold identified as optimal trade-off between model precision and recall when aiming to reduce false-negatives as much as possible.

Threshold [%]	TN [%]	FP [%]	FN [%]	TP [%]	TPR	TNR
0	0	86	0	14	1.00	0.00
10	62	24	1	13	0.93	0.72
20	75	12	3	11	0.79	0.86
30	**80**	**6**	**4**	**10**	**0.71**	**0.93**
40	84	2	6	8	0.57	0.98
50	85	1	7	7	0.50	0.99
60	86	0	9	5	0.36	1.00
70	86	0	11	3	0.21	1.00
80	86	0	12	2	0.14	1.00
90	86	0	14	0	0.00	1.00
100	86	0	14	0	0.00	1.00

TN - true negatives FP - false positives FN - false negatives
TP - true positives TPR - true positive rate TNR - true negative rate

2.4.1 Clinical needs

The development of a novel decision support system in NSTI is of special clinical relevance, as scoring systems currently in use have several disadvantages compared to a CDSS. Firstly, scoring systems are developed for a specific clinical endpoint, such as AKI [151], mortality [168], or septic shock [169]. Secondly, they are retrospectively designed, static, and incapable of easily incorporating novel patient information. Furthermore, they have limited flexibility regarding missing input data. Comparable decision support tools, such as the *AKIpredictor*, performed as as good as physicians, with ROC-AUC scores of 0.94 (95% CI, 0.89–0.98) (physician) versus 0.89 (95% CI, 0.82–0.97) (*AKIpredictor*) [170]. The interviews with various clinicians of different specialities provide a starting point of relevant clinical questions, on which the CDSS can be based. Highly relevant questions were identified in various domains, as early as upon presentation to the GP. However, although questions in all phases would ideally be supported, the currently available INFECT data set limits the possibilities to perform predictions for questions in the pre-ICU phases, mainly because of the lack of sufficient pre-hospital data as well as the lack of non-NSTI patients for diagnostic predictions. Therefore, the initial CDSS will be designed for use in the ICU, and include the most relevant questions in the ICU phase. When adjacent data sets become available, this could be expanded in the future.

A

Variable	Availability	Set name
Age	General	GENERAL
Lowest systolic BP (BL)	vital parameters	
Highest pulse (BL)	10 minutes	
Lactate (BL)	Blood 1 hour	BLOOD
Base excess (BL)		
INR (BL)		
Bicarbonate (BL)		
pH (BL)		
Bilirubin (BL)		
CRP (BL)		
Creatinine (BL)		
Platelets (BL)		
Carbamide (BL)		
Noradrenaline (BL)	Treatment 1 hour	TREATMENT
Total urine (BL)	After 24 hour	24 HOURS
Total fluids (BL)		

B

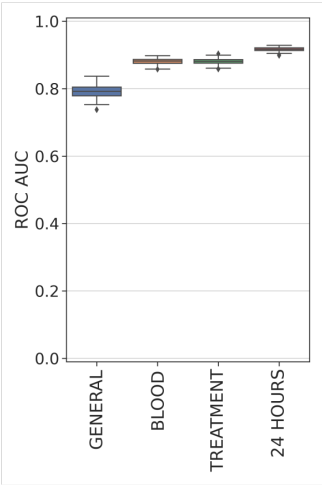


Figure 2.6: **Taking variable availability into account.** (A) Variables included in subsets of BL model. Time points indicate the approximate time of availability in the ICU. (B) Prediction performances of models trained on selected variable subsets.

2.4.2 Proof of concept

Our results clearly demonstrate that the developed system proves to be proficient in estimating patient mortality risks early on and with performances that are comparable (and in some cases better) than scoring systems currently in use. Additionally observed benefits of our CDSS include flexibility and robustness towards changes in the input parameters. The variables selected to be relevant for mortality predictions are diverse, ranging from demographics and vital measurements to information obtainable only after a certain time period spent in the ICU. However, all of the selected variables are well-known parameters when assessing the vital state of patients, especially in the ICU. Therefore, many of the identified variables can also be found in established scoring systems like the SAPS II (age, pulse, systolic blood pressure, bicarbonate, carbamide)[165], SOFA score (platelets, bilirubin, creatinine, urine output, vasopressor requirement)[147], or APACHE IV score (pH) [149]. Also previously described as associated with mortality were lactate [171, 172], base excess [173], C-reactive protein [174], and INR [175]. Since fluid requirement increases depending on sepsis severity, this predictor was expected as well. More computationally-oriented publications comparable to the work carried out here have reported similar variables as informative, although their ranking of importance differs [155, 168, 176, 177]. The fact

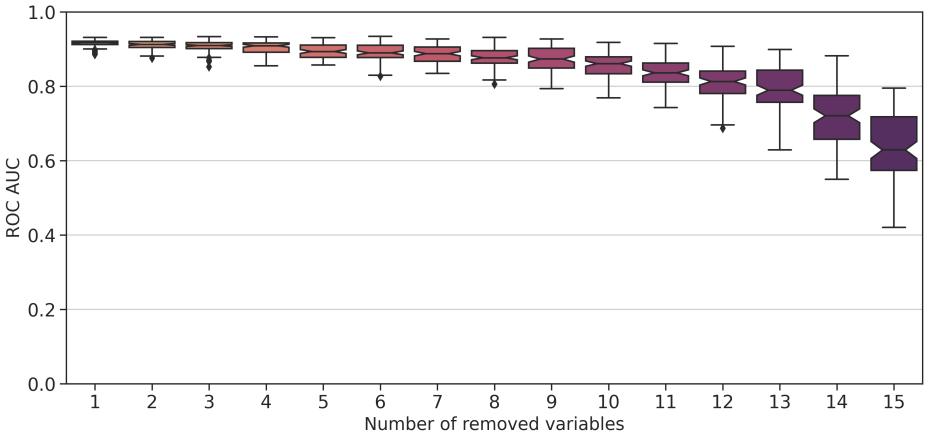


Figure 2.7: **Assessment of model stability through iterative removal of random variables.** 24 HOUR set with a total number of 16 variables. Displayed as average ROC-AUC score over 100 iterations. Notches represent 95% confidence interval around the median.

that all selected variables were previously found to be associated with risk of mortality in sepsis patients validate the findings from this study. It is evident that the selected variables do not include information unique to NSTI, such as the type of causative micro-organism, anatomical location affected, or surgical findings. This is likely due to the fact that mortality of these patients in the ICU is primarily due to sepsis, making it logical that identified predictors are connected to the systemic disease rather than the local characteristics of NSTI. Estimation of NSTI specific endpoints, such as wound size or need for amputation, may yield a more specialised set of predictive variables.

2.4.3 Practicality

The final CDSS envisioned should not merely deliver a binary prediction of patient outcomes, but rather give information on the likelihood of an event, which can subsequently be interpreted by clinicians. When examining the calculated likelihood of death for each patient, the CDSS notably seems much more capable of identifying survivors than patients at risk. This pronounced difference might be ground in the heavy imbalance observed in the data set, potentially favoring the identification of alive patients. The observed clustering of deceased patients during the first week is explainable from a clinical perspective, since patients with refractory septic shock will often die in the first day after admission. Since those early deaths represent the majority of those who died during admission, it is unsurprising that most predictors selected are related to septic shock, and may therefore

Table 2.2: **Overview on selected surrogate variables and their respective Pearson correlation to the original variable.** Best performing surrogates are highlighted in bold.

Original	Surrogate	Correlation
Age	-	-
Lowest systolic BP (BL)	Skin bullae (preop)	-0.17
Highest pulse (BL)	Potassium (BL)	0.27
Lactate (BL)	Potassium (BL)	0.26
Base excess (BL)	Creatinine (preop)	-0.33
INR (BL)	Chronic liver disease	0.29
Bicarbonate (BL)	Creatinine (preop)	-0.36
pH (BL)	Creatinine (preop)	-0.37
Bilirubin (BL)	Chronic liver disease	0.26
CRP (BL)	**CRP (preop)**	**0.67**
Creatinine (BL)	**Creatinine (preop)**	**0.88**
Platelets (BL)	WBC (BL)	0.35
Carbamide (BL)	**Creatinine (preop)**	**0.63**
Noradrenaline (BL)	Potassium (BL)	0.22
Total urine (BL)	Creatinine (preop)	-0.25
Total fluids (BL)	Corticosteroid use	0.22

lead to the most accurate predictions for early deaths.
The exceptional stability of our CDSS is of high clinical relevance, as missing measurements are frequent in a typical clinical setting. Strategies of handling missing data when working with the existing scoring systems include imputation strategies such as i) replacement by a previous measurement or ii) assuming the value to be in the normal physiological range. Although easy to use, both of these approaches may introduce erroneous data points potentially biasing predictions and should thus be applied with care. Not all variables possess equal explanatory power, leading to more and less favorable scenarios of data missingness. Suggestions of alternative measurements could mitigate the effect of missing variables, especially of those with high predictive power. Our results suggest that including alternative variables can be successfully implemented in the case of well-suited surrogates (e.g. with correlation > 0.5) and is of special importance when multiple measurements are missing.

2.4.4 Strengths and limitations of our study

By engaging the clinical specialists upfront through interviews, we were able to not only clearly identify clinical needs, but also rank them according to relevance. Although no

classic qualitative approach was applied and the ranking system used is not a validated method, we believe the obtained overview is sufficient to act as a starting point for the design of a CDSS. For the development of a mortality prediction model, we used the largest NSTI cohort available as of today, which is a major strength. However, the model development still suffered from small sample sizes with heavy imbalance, potentially limiting validity and generalizability. Despite using widely approved imputation strategies to account for missing entries in the data set, the need for data imputation remains a limitation in the pre-processing pipeline. Although the performance of our CDSS is yet to be externally validated, we have taken care of deploying robust internal validation techniques during variable selection and model development phases. The final framework uses a small number of predictors that are readily available in most ICUs, which facilitates international use. The possibility of providing various input variables allows this CDSS to be used at different stages throughout the treatment process. Despite the potential benefits of a CDSS it must be stressed that there is an inherent risk for misuse, as with any tools of this kind. Special care must be taken when relating probabilities to outcomes, as e.g. the intuitive classification threshold of 50% does not translate to a 50% risk of mortality. Proper application and consistent monitoring of results will be essential in fostering trust in a DSS utilizing machine learning algorithms.

2.4.5 Future directions

To successfully deploy the comprehensive multiple-endpoint CDSS envisioned, future efforts will cover several areas including data management, model optimization, framework extension, and deployment. Firstly, the ongoing efforts to establish a data management system underlying the predictive models will be intensified. This will facilitate easy addition of novel patient information, paving the way for the models to 'learn as they grow'. Secondly, the predictive models developed in the course of this study will be i) validated using external cohorts ii) extended to cover a wider array of clinical questions identified in the interviews. The integration of more diverse data types will be aspired (e.g. -omics data) to refine patient stratification and deduce underlying biological disease mechanisms. The lack of NSTI-specific information needed to estimate the risk of mortality provides the opportunity of testing the developed framework on a more general sepsis cohort, thereby broadening its field of application. Simultaneously, however, the inclusion of more NSTI-specific variables will enable the creation of a specialized framework addressing health endpoints unique to NSTI, such as the prediction of suspected microbiological species. Lastly, we seek to deploy models in the most user-friendly way by developing (free) applications tailored to end-user needs (including web, phone, and tablet apps).

2.5 Conclusion

In summary, our study lays the foundation of a comprehensive CDSS for NSTI patients. To the best of our knowledge, we have for the first time provided a qualitative assessment of the clinically relevant questions in NSTI patient care. By intertwining clinical and bioinformatic expertise, we have developed a tool proficient in predicting 30-day mortality for patients with NSTI admitted to the ICU. The possibility for users to adapt the input variables without severely affecting model performance is a major advantage compared to other clinical scoring systems currently in use, as it is the possibility of continuously updating the predictors as patients are included. Furthermore, the framework itself can be easily expanded to other health endpoints related to NSTI diagnosis and treatment such as suspected microbiological species, risk of septic shock, or risk of acute kidney injury, thus creating a universal tool for improving NSTI care and outcomes.

2.5 Supplementary Notes

All Supplementary Material, including figures and tables, are available under: https://doi.org/10.1016/j.ijmedinf.2022.104878

2.5.1 Acknowledgements

The authors are grateful to all members of the INFECT, PerAID and PerMIT collaborations, as extensive discussion within the consortium have greatly contributed to the finalisation of this manuscript. Beside named authors in this article, the INFECT Study Group includes: Michael Nekludov, Ylva Karlsson, Per Arnell, Morten Hedetoft, Marco B. Hansen, Peter Polzik, Daniel Bidstrup, Nina F. Bærnthsen, Gladis H. Frendø, Erik C. Jansen, Lærke B. Madsen, Rasmus B. Müller, Emilie M. J. Pedersen, Marie W. Petersen, Frederikke Ravn, Isabel F. G. Smidt-Nielsen, Anna M. Wahl, Sandra Wulffeld, Sara Aronsson, Anders Rosemar, Joakim Trogen, Trond Bruun, Torbjørn Nedrebø, Oddvar Oppegaard, Eivind Rath and Marianne Sævik. The PerAID/PerMIT Study groups, besides named authors, include: Mattias Svensson, Kristoffer Stralin, Trond Bruun, Oddvar Oppegaard, Knut Anders Mosevoll, Jan Kristian Damas, Paul van Zuijlen, Laura M. Palma Medina, Lorna Morris, and Marco Anteghini.

2.5.2 Funding

The study was supported by 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, 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); the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860895 TranSYS.

Chapter 3

COVID-19 and cholesterol biosynthesis: Towards innovative decision support systems

Eva Kočar[†], **Sonja Katz**[‡], Žiga Pušnik[‡], Petra Bogovič, Gabriele Turel, Cene Skubic, Tadeja Režen, Franc Strle, Vitor A.P. Martins dos Santos, Miha Mraz, Miha Moškon[‡], Damjana Rozman

[†]authors contributed to the study design, data collection, and performing of wet-lab experiments (section 3.2.2)

[‡]authors contributed to the implementation of computational models (section 3.2.3)

Published in: iScience. 2023 Oct 20;26(10)
DOI: 10.1016/j.isci.2023.107799

Abstract

With COVID-19 becoming endemic, there is a continuing need to find biomarkers characterizing the disease and aiding in patient stratification. We studied the relation between COVID-19 and cholesterol biosynthesis by comparing 10 intermediates of cholesterol biosynthesis during the hospitalization of 164 patients (admission, disease deterioration, discharge) admitted to the University Medical Center of Ljubljana. The concentrations of zymosterol, 24-dehydrolathosterol, desmosterol, and zymostenol were significantly altered in COVID-19 patients. We further developed a predictive model for disease severity based on clinical parameters alone and their combination with a subset of sterols. Our machine learning models applying 8 clinical parameters predicted disease severity with excellent accuracy (AUC = 0.96), showing substantial improvement over current clinical risk scores. After including sterols, model performance remained better than COVID-GRAM. This is the first study to examine cholesterol biosynthesis during COVID-19 and shows that a subset of cholesterol-related sterols is associated with the severity of COVID-19.

3.1 Introduction

COVID-19 is still very much present and, according to epidemiologists, its causative agent SARS-CoV-2 is likely to become endemic with new variants recurring seasonally [178]. Therefore, searching for new biomarkers for disease course and outcome prediction remains of high importance.

Viral infections can trigger changes in the host organism's lipid profile, which could serve as a biomarker. In fact, dyslipidemia seems to be a hallmark of viral infections, as previously shown in hepatitis C virus (HCV), human immunodeficiency virus (HIV), and dengue virus infection [179]. In 2020, Hu et al. [180] were among the first to report an altered lipid profile in COVID-19 patients. Serum total cholesterol as well as HDL- and LDL-cholesterol levels were significantly lowered in patients suffering from COVID-19. Reports about dyslipidemia linked to SARS-CoV-2 infection from other groups soon followed [181–194], but findings were not always concordant. Furthermore, increased serum levels of the liver enzymes alkaline phosphatase (ALP), alanine aminotransferase (ALT), and aspartate aminotransferase (AST) seen in COVID-19 patients were suggested to be explained by liver damage as a result of the infection [179, 182]. Chen et al. [188] showed that liver-specific proteins that regulate sterol and cholesterol transport were downregulated in COVID-19 patients. Although altered blood cholesterol levels have been reported in patients suffering from COVID-19, the effect of COVID-19 on intracellular biosynthesis of cholesterol remains to be determined.

Besides being a fundamental lipid component of vertebrate cell membranes, primarily as a lipid building block of ordered membrane micro-domains – lipid rafts [179, 195, 196] – cholesterol also modulates their permeability, signaling, and transport. Furthermore, it can be integrated into lipoproteins, and stored in lipid droplets and cholesteryl esters [179, 197]. Importantly, its metabolic pathways branch out in several ways, resulting in physiologically important, active compounds (e.g., bile acids, oxysterols, steroid and glucocorticoid hormones, vitamin D, coenzyme Q) [179, 197–199]. As cholesterol is an important molecule with versatile functions in numerous physiological processes and an excess of non-esterified cholesterol may potentially be toxic, maintaining its homeostasis is pivotal [179, 200]. Cholesterol biosynthesis is a tightly regulated housekeeping pathway and takes place mainly in the liver. The pre-squalene part of *de novo* cholesterol biosynthesis starts with acetyl-CoA and terminates with the enzymatic conversion of farnesyl-PP to squalene (Figure 3.1). A more detailed description of the pre-squalene pathway is given elsewhere [201]. Conversion of squalene to lanosterol is the link between the pre- and post-squalene parts of the biosynthesis. Lanosterol is converted through a series of enzy-

matic reactions to cholesterol via the Bloch and/or Kandutsch-Russell (K-R) pathway [199, 201, 202]. Both pathways are enzymatically identical, except for the first and the last steps (Figure 3.1). In the Bloch pathway, CYP51A1 catalyzes the conversion of lanosterol to FF-MAS, while in the K-R pathway, sterol-Δ24-reductase (DHCR24) catalyzes the same sterol intermediate to 24,25-dihydrolanosterol. In the last enzymatic step of the Bloch pathway desmosterol is converted to cholesterol by DHCR24, while the final reaction of the K-R branch is the conversion of 7-dehydrocholesterol to cholesterol by DHCR7. Since all sterol intermediates (referred to as sterol intermediates or sterols throughout the manuscript) from lano- to desmosterol in the Bloch branch contain Δ24 double bonds, DHCR24 can in principle metabolize any of them, and both branches, therefore, intertwine. Nonetheless, an in vitro study showed that 24-dehydrolathosterol is the most plausible switching point between both branches, indicating that cholesterol biosynthesis preferentially starts via Bloch and later shifts to the K-R pathway [203]. Depending on tissue type, one or an interplay of both pathways dominates *de novo* cholesterol biosynthesis.

The aim of this study was to evaluate the effects of COVID-19 on intracellular cholesterol biosynthesis, to develop a predictive model for the severity of COVID-19 course based on simple clinical parameters obtained at hospital admission using machine learning models, and to investigate the potential benefits of measuring metabolic pathways for disease monitoring. Including metabolic biomarkers into clinical diagnostics is attracting growing interest, emphasizing the importance of investigating their potential use in this context. Previous research has focused on altered blood cholesterol levels in COVID-19 patients, but detailed knowledge of endogenous cholesterol biosynthesis in patients with SARS-CoV-2 infection is scarce. In addition, existing prognostic models have limitations and lack reproducibility across different populations. Thus, we sought to identify readily available clinical variables and reliable methods to predict disease severity beyond the established COVID-GRAM risk score, as well to investigate biomarker potential of sterols. We performed targeted lipidomics to monitor changes in blood sterol intermediates – indicators of *de novo* intracellular cholesterol biosynthesis – in hospitalized COVID-19 patients. In addition, machine learning techniques were used to retrospectively estimate disease severity based on clinical parameters alone and also to investigate the biomarker potential of cholesterol-related sterols measured at hospital admission.

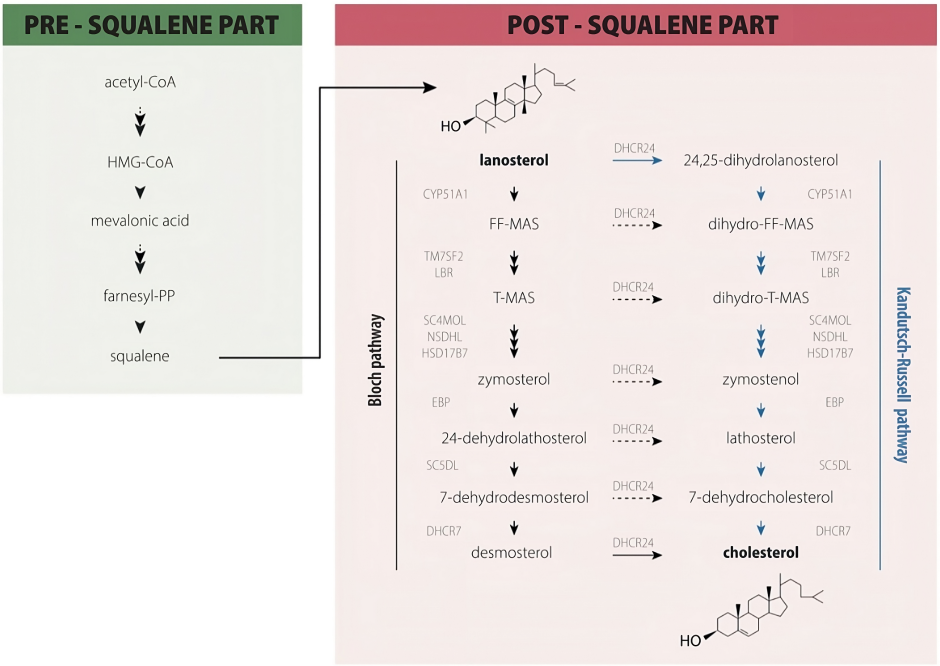


Figure 3.1: **The cholesterol biosynthesis pathway**. The cholesterol biosynthesis pathway is divided into pre- and post-squalene parts. In the pre-squalene part, acetyl-CoA is converted to squalene through a series of enzymatic reactions. The conversion of squalene to lanosterol represents a linking point between the two parts of the biosynthesis. After lanosterol, the biosynthesis branches into the Bloch and Kandutsch-Russell (K-R) pathways. In the Bloch pathway, lanosterol is converted to FF-MAS by CYP51A1, while in the K-R pathway DHCR24 converts it to 24,25-dihydrolanosterol. The final reaction in the Bloch pathway is a conversion of desmosterol to cholesterol by DHCR24, whilst in the K-R pathway DHCR7 catalyzes the conversion of 7-dehydrocholesterol to cholesterol. Both pathways can intertwine but data show that 24-dehydrolathosterol is the most plausible substrate where switching occurs, indicating that cholesterol biosynthesis preferentially starts via the Bloch and later shifts to the K-R pathway. Enzymes catalyzing each reaction are shown in gray. Sterol chemical names are listed in Table S1.

3.2 Results

3.2.1 Cohort description

164 adult patients admitted to the Department of Infectious Diseases of the University Medical Center Ljubljana (Slovenia) from July 2020 to July 2021 suffering from a severe course of COVID-19 were enrolled in this study. Their basic clinical characteristics are shown in Tables 3.1 and S2.

3.2.2 COVID-19 impact on *de novo* intracellular cholesterol biosynthesis

As the concentration of blood cholesterol depends on different factors, e.g., *de novo* biosynthesis and dietary cholesterol, the rate of *de novo* intracellular cholesterol biosynthesis can only be estimated according to the presence of sterol intermediates. We used liquid chromatography with tandem mass spectrometry (LC-MS/MS)-based targeted lipidomics to provide insight into cholesterol intermediates during COVID-19. Ten sterols from the post-squalene part of cholesterol biosynthesis were measured in serum samples at three different time points during hospitalization of 62 COVID-19 patients, i.e., lanosterol, 24,25-dihydrolanosterol, T-MAS, dihydro-T-MAS, zymosterol, zymostenol, 24-dehydrolathosterol, lathosterol, desmosterol, and cholesterol. Samples were collected upon admission to the hospital care (T1), in case of severe deterioration or in the middle of treatment (T2), and upon discharge (T3). An additional 102 COVID-19 patients serum samples were collected only at T1 and sterol intermediates were measured. Concentrations of sterol intermediates at all time points are shown in Table 3.2. Statistical significance was tested using the nonparametric Friedman test (Figures 3.2A and 3.2B) comparing three time points of sterol concentrations. For multiple comparisons, the adjusted p-values were determined using Dunn's test. Results show more significant alterations in cholesterol biosynthesis in patients with severe (Figure 3.2B), compared to those with mild disease course (Figure 3.2A). In the latter, statistically significant changes during the course of the disease were found in the concentrations of 24-dehydrolathosterol (T1 vs. T2), zymostenol (T1 vs.T2), and cholesterol (T1 vs. T2), whereas in patients with severe course of COVID-19 statistically significant changes were observed in zymosterol (T1 vs. T2, T1 vs. T3), 24-dehydrolathosterol (T1 vs. T3, T2 vs. T3), desmosterol (T1 vs. T3, T2 vs. T3), zymostenol (T1 vs. T2) and finally cholesterol (T1 vs. T3), most of them being representatives of the Bloch pathway. Findings are shown in Figures 3.2A and 3.2B, Tables 3.2 and S3.

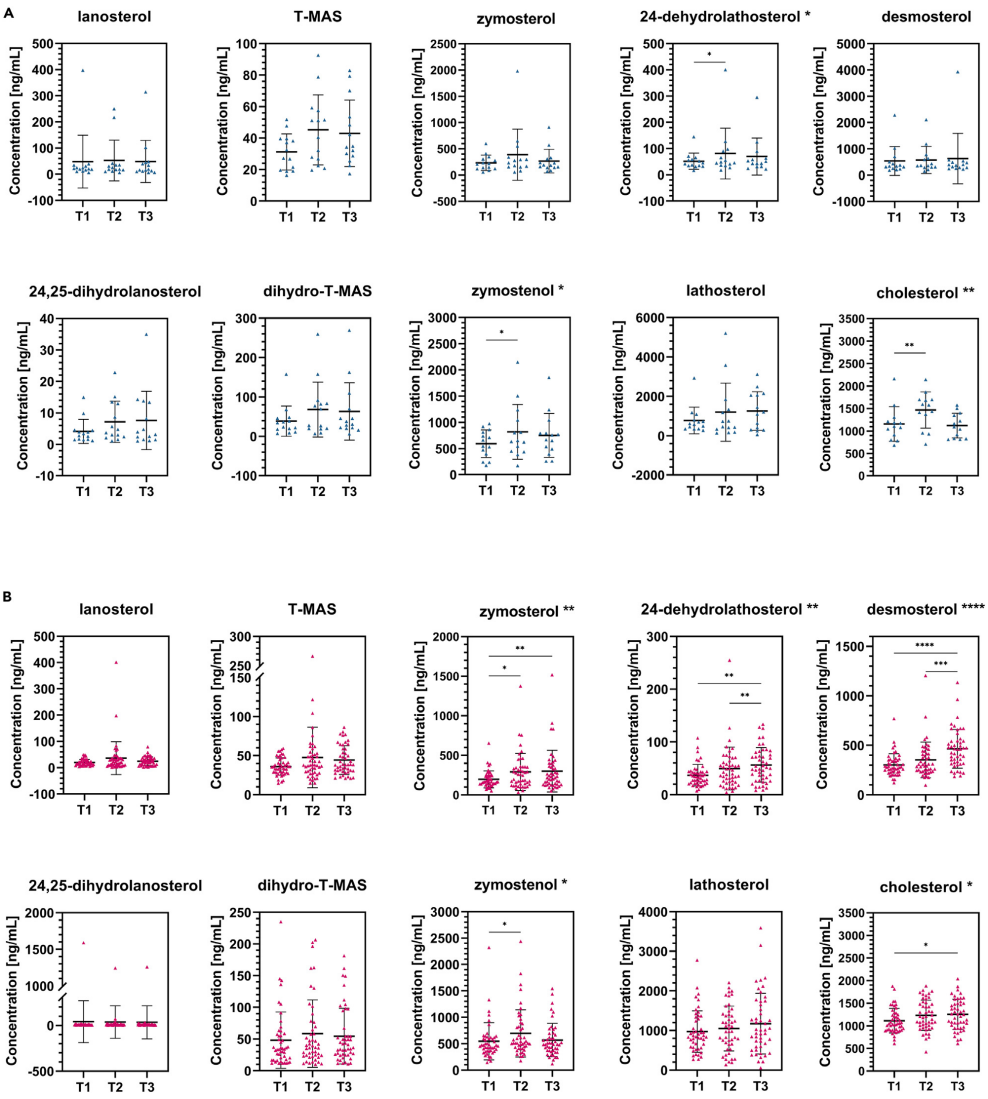


Figure 3.2: **Changes in serum lipid profile during COVID-19**. (A and B) Changes in serum lipid profile during COVID-19 in patients with (A) mild and (B) severe course of the disease tested with the nonparametric Friedman's test followed by Dunn's post hoc test for multiple comparisons testing. Statistical significance of Friedman's and Dunn's tests is indicated beside the sterol name and inside the graph, respectively. Data are represented as mean G SD. See also Table S3. SD, standard deviation. * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001, **** ≤ 0.0001

Table 3.1: **Basic demographic, clinical, and biochemical parameters of hospitalized COVID-19 patients**. The comorbidities section contains the presence of diabetes, hyperlipidemia, thyroid disease, arterial hypertension, heart failure, chronic disease of lung, liver, and kidney, rheumatic disease, active malignant disease, and/or transplantation. See also Table S2. Disease category data are number (%); other values are given as mean ± SD. BMI, body mass index; CRP, C-reactive protein; LDH, lactate dehydrogenase; SD, standard deviation.

Number of patients	164
Age	61.0 ± 14.0
BMI	31.9 ± 6.7
Disease category	
Mild	7 (4.27%)
Moderate	23 (14.02%)
Severe	100 (60.98%)
Critical	34 (20.73%)
Comorbidities	
yes	125
no	39
Other parameters	
*Reason for admission**	need for oxygen: 119 subjective feeling of difficulty breathing: 17 worsening of the underlying diseases: 6 other: 23
*Oxygen saturation measured with or without oxygen supplementation**	oxygen: 95 air: 69
*Oxygen saturation [%]**	94.0 ± 3.0
*CRP [mg/L]**	104.4 ± 75.1
*Ferritin [μg/L]**	1008.7 ± 839.4
*LDH [μkat/L]**	5.72 ± 1.99
*X-ray – lung abnormalities**	yes: 140 no: 22
*X-ray – lung opacities locations**	unilateral: 16 bilateral: 121
In-hospital outcome	
Survival	159
Deaths	5

*At admission to hospital.

Table 3.2: **Concentration of sterol intermediates in patients with mild (N = 14) and severe (N = 48) COVID-19 measured at hospital admission.** Data are represented as mean ± SD. See also Table S3. SD, standard deviation.

	Concentration [ng/mL]					
	T1		T2		T3	
Sterol name	*Mild*	*Severe*	*Mild*	*Severe*	*Mild*	*Severe*
lanosterol	48.34 ± 100.98	19.95 ± 11.77	52.65 ± 77.73	36.21 ± 62.40	48.76 ± 80.54	24.42 ± 16.60
24,25-dihydrolanosterol	4.15 ± 3.83	40.46 ± 228.38	7.18 ± 6.54	36.09 ± 178.55	7.63 ± 9.27	33.92 ± 180.78
T-MAS	31.18 ± 11.46	36.13 ± 11.16	45.17 ± 22.25	47.65 ± 38.81	43.02 ± 21.14	44.32 ± 18.36
dihydro-T-MAS	38.83 ± 38.33	48.17 ± 44.41	68.10 ± 69.70	58.52 ± 53.27	63.56 ± 72.67	54.53 ± 43.83
zymosterol	234.29 ± 146.37	196.70 ± 107.56	387.69 ± 487.68	291.07 ± 234.56	266.49 ± 224.34	300.75 ± 262.96
zymostenol	593.27 ± 263.01	546.95 ± 352.53	819.32 ± 521.05	696.03 ± 446.04	752.53 ± 419.62	569.46 ± 314.30
24-dehydrolathosterol	51.84 ± 30.73	36.94 ± 20.31	81.06 ± 96.69	49.50 ± 40.53	69.65 ± 70.39	56.50 ± 32.96
lathosterol	781.71 ± 677.21	974.93 ± 524.65	1200.01 ± 1475.55	1050.77 ± 565.37	1258.76 ± 979.63	1171.96 ± 764.83
desmosterol	538.41 ± 548.88	302.72 ± 112.58	577.32 ± 516.68	352.64 ± 180.33	630.24 ± 957.82	465.34 ± 195.47
cholesterol	1156.96 ± 385.43	1112.98 ± 271.54	1442.01 ± 398.03	1233.75 ± 335.29	1107.95 ± 267.97	1253.15 ± 336.29

3.2.3 Developing machine learning models for COVID-19 course prediction

The specific aims of the study were (1) to predict disease severity based on clinical parameters using machine learning models better than the currently established COVID-19 clinical risk score (i.e., COVID-GRAM [204]), and (2) to evaluate whether the inclusion of sterol intermediates could strengthen the prediction performance.

The cohort utilized in this study recorded more than 70 clinical variables upon admission and measured concentrations of 10 sterols at T1. Using all clinical variables to estimate disease severity would lead to overfitting of machine learning models, resulting in poor generalizability. Therefore, to identify a smaller subset of meaningful variables, we conducted an unsupervised variable selection to identify important predictors for disease severity among clinical, as well as sterol intermediates measurements. This yielded a total of 8 clinical and 4 sterol variables (Table S4), which included the reason for hospital admission, information on how oxygen saturation was measured and its level, conclusions drawn from X-ray analyses, and concentrations of ferritin, LDH, CRP, 24,25-dihydrolanosterol, zymostenol, 24-dehydrolathosterol, and desmosterol. An overview of these variables can be found in Table 3.3, with a more detailed description in Table S5.

To evaluate the predictive power of the selected clinical variables, T1 sterols, and their combination, we first trained eight different machine learning models on each scenario using leave-one-out-cross-validation. Their performances were compared with several metrics and the best performing model selected based on highest scores across all metrics, model simplicity, and interpretability (Figure S1; Table S6). The best performing models of each scenario were subsequently compared to each other (Figure 3.3A; Table 3.4). A model trained on clinical variables alone showed excellent predictive power ("clinical",

Table 3.3: **Summary overview on variables, measured at admission to hospital, contained in each variable set (clinical, T1 sterols, clinical + T1 sterols**. See also Tables S4 and S5. CRP, C-reactive protein; LDH, lactate dehydrogenase.

Variable	Scenario	
Reason for admission	clinical	clinical + T1 sterols
Oxygen saturation measured with or without oxygen supplementation		
Oxygen saturation [%]		
CRP		
Ferritin		
LDH		
X-ray – lung abnormalities		
X-ray – lung opacities locations		
24,25-dihydrolanosterol	T1 sterols	
zymostenol		
24-dehydrolathosterol		
desmosterol		

AUC = 0.96). Based on T1 sterol measurements alone, disease severity can be estimated with moderate confidence ("T1 sterols", AUC = 0.66). The performance of the model remained excellent when both groups of variables were combined ("clinical + T1 sterols", AUC = 0.95).

To gain more insight into the decision making of each model, feature importance for each scenario was evaluated using a permutation-based feature importance method (see STAR methods, classification models), revealing sets of features with the highest importance in the prediction of COVID-19 severity (Figure 3.3B). We observe that although variables were selected in a data-driven manner, not all of them display equal importance during the predictive task. While the primary reason for hospital admission ("reason for admission"), CRP, and the measurement method of oxygen saturation ("Oxygen saturation measured with or without oxygen supplementation") remain highly important across scenarios, we see a shift in the importance of, e.g., X-ray measurements or 24,25-dihydrolanosterol upon combining clinical and T1 sterols information. This may indicate that the selected sterols correlate with the respective clinical measurements, resulting in an overlap of information (Figure S2).

In order to put our model performances in the context of established clinical practices, we compared them to the COVID-GRAM risk score, as well as a "clinical baseline" model we developed using three variables commonly used to stratify patients into disease severity groups, namely concentrations of LDH, ferritin, and CRP upon admission (Figure 3.3C). It is evident that the "clinical" as well as combined "clinical + T1 sterols" models greatly outperform the COVID-GRAM and clinical baseline estimates. We can also observe that

while the clinical baseline model can predict disease severity fairly accurately, the inclusion of only a handful of additional variables would greatly benefit model accuracy. Sterol measurements alone show moderate performance similar to COVID-GRAM risk estimates. In general, the validity of COVID-GRAM for this cohort is to be questioned, as although it uses a similar set of variables to the "clinical" and "clinical baseline" model it seems to be unable to delineate patient trajectories.

Table 3.4: **Evaluation metrics for binary classification for best performing classifiers in each scenario**. See also Table S6. AUC, area under the ROC curve; f1, F1-score; GNB, Gaussian Naive Bayes; MCL, Majority Classifier; RFC, Random Forest.

Dataset	Classifier	Performance metric				
		precision	recall	f1	accuracy	AUC
clinical	GNB	0.976	0.910	0.942	0.909	0.955
T1 sterols	RFC	0.849	0.963	0.902	0.829	0.664
clinical + T1 sterols	RFC	0.926	0.940	0.933	0.890	0.950
	MCL	0.817	1.000	0.899	0.817	0.500

3.3 Discussion

3.3.1 Viral infections lead to changes in the lipid metabolism of the host

It is well known that host undergoes a number of physiological changes during viral infection. Studies have previously demonstrated significant alterations in lipid metabolism driven by bacterial and viral infection [179, 180, 205–211], including SARS-CoV-2 [180, 182, 212–225]. The degree of hypolipidemia was shown to inversely correlate with disease severity and fatality rate in COVID-19 patients [185, 213, 221, 226, 227]. Additionally, a few studies reported progressive changes in levels of different phospholipids [220] and sphingolipids [224], suggesting lipidome signature as a putative biomarker of disease severity and outcome.

A handful of papers have reported viral infection affecting levels of sterol intermediates. Mercorelli et al. [228] have recently shown that human cytomegalovirus (HCMV) increased CYP51 gene expression in the U-373 MG cell line, whose protein product is in charge of the enzymatic conversion of lanosterol to FF-MAS. Furthermore, CYP51 expression is also induced by HIV-1 Nef protein in order to upregulate cholesterol biosynthesis and its transport to lipid rafts, resulting in increased virion infectivity [229]. In contrast,

significantly decreased levels of serum lathosterol were seen in HCV genotype 3 patients [230] while metabolite profiling of JFH1 cell culture infected with HCV demonstrated an approximately 10-fold accumulation of desmosterol [231, 232]. However, to our knowledge there is no report that provides a detailed characterization of endogenous cholesterol biosynthesis during viral infection. The results of our study show that in patients suffering from severe COVID-19, the Bloch pathway (three of five sterol intermediates – zymosterol, 24-dehydrolathosterol, desmosterol) and, to a lesser extent, the K-R pathway (one of four sterol intermediates – zymostenol) of endogenous cholesterol biosynthesis are significantly impaired (Figures 3.1 and 3.2B). In patients with mild/moderate disease course only 24-dehydrolathosterol from the Bloch and zymostenol from the K-R pathway were significantly altered (Figures 3.1 and 3.2A). However, serum concentrations of cholesterol were significantly elevated in both patient groups (Figures 3.2A and 3.2B). At hospital discharge, an increase in the concentrations of T-MAS, zymosterol, zymostenol, 24-dehydrolathosterol, lathosterol, and desmosterol was noted, indicating that their levels started to recover. This is consistent with previous reports on other biochemical parameters [182, 214, 233, 234]. However, we cannot say with certainty that they reached their basal level. Accordingly, some studies report that alterations of some metabolic parameters persist in recovered patients for a longer period of time, suggesting a long-term systemic effect on the hosts' metabolism following SARS-CoV-2 infection [225, 235–237]. Different durations and stages of illness at admission to hospital, and the fact that T2 samples were collected either in the case of severe deterioration or in the middle of hospitalization might explain higher standard deviations seen in several sterol intermediates (i.e., 24,25-dihydrolanosterol, T-MAS, zymostenol, and lathosterol; Figures 3.2A and 3.2B). However, the broad distribution of sterol concentrations in T2 could also indicate a different course of the disease between individual patients.

Plasma levels of sterol intermediates reflect liver function. Liver injury seen in COVID-19 patients, especially in those with severe or critical course of the disease, may be a result of a variety of mechanisms including direct viral damage, systemic inflammation, immune injury, hypoxia and ischemia, drug-induced liver injury as well as worsening of underlying liver diseases [238–240].

3.3.2 Data-driven variable selection identifies a set of clinical variables highly predictive of disease severity

Patients suffering from severe COVID-19 may experience rapid deterioration and admission to the intensive care unit, which represents, on the one hand, a threat to the patient's life and, on the other, a considerable burden on the medical system, as experienced during

the first epidemic waves [241]. Timely recognition and correct prognosis of the disease are essential for the optimal use of available resources. Therefore, we investigated whether we could better predict the development of severe COVID-19 in our patient cohort using machine learning models based on clinical parameters measured at hospital admission than using the currently available clinical risk score COVID-GRAM. In the second part, we aimed to delineate a possible biomarker potential of sterols in predicting disease severity. Our analyses revealed that a set of 8 clinical measurements is sufficient to predict disease severity with high accuracy (Tables 3.3 and S5). Among the most important clinical variables were those found to be the reason for admission to hospital, such as the need for oxygen treatment or difficulties breathing, and "oxygen saturation measured with or without oxygen supplementation", which detailed whether patients at the time of saturation measurement evidently need supplemental oxygen (and were therefore receiving oxygen therapy) or not (they were breathing normal air). These findings are not surprising, as dyspnea has been reported as an important factor determining COVID-19 course [242]. Hentsch et al. [243] showed that perceived breathlessness usually occurs in an advanced stage of the disease, which suggests that the reason for admission might be an indicator on how far the disease has progressed prior to hospital admission. Also, several biochemical parameters, namely concentrations of LDH, ferritin, and CRP, were deemed important for severity predictions. Increased concentration of CRP, a type I acute phase response protein synthetized in the liver and regulated by the pro-inflammatory cytokines IL-6, IL-1, and TNF-α, correlates with disease severity and predicts a need for ICU treatment, as well as mechanical ventilation. Patients with a critical course of COVID-19 also show elevated levels of ferritin, D-dimer, and lactate dehydrogenase (LDH), which are associated with poor prognosis and outcome [242, 244, 245]. Our analyses further found chest X-ray information to be relevant, which is supported by recent publications showing the usefulness of X-rays in risk stratification for clinical worsening and prediction of fatality rate in COVID-19 patients [246, 247].

In addition, we found that a subset of four sterols was associated with disease severity, namely 24,25-dihydrolanosterol, zymostenol, 24-dehydrolathosterol, and desmosterol (Table 3.3). The latter three were also significantly altered during the course of the disease (Figures 3.2A and 3.2B). We believe that this is not a coincidence, since oxygen is needed for several enzymatic reactions in cholesterol biosynthesis and COVID-19 patients suffer from hypoxemia [203]. Three molecules of oxygen are needed for conversion of lanosterol to FF-MAS (or 24,25-dihydrolanosterol to dihydro-FF-MAS) by CYP51A1 and for conversion of T-MAS to zymosterol (or dihydro-T-MAS to zymostenol) by SC4MOL, while one molecule of oxygen is required for transformation of 24-dehydrolathosterol to

7-dehydrodesmosterol (not measured within this study) and further to desmosterol by SC5DL and DHCR7, respectively (Figure 3.1). The most significant change during COVID-19 was seen in desmosterol concentrations in severe COVID-19 patients (Figure 3.2B) which is not surprising as it is the last precursor before cholesterol and the effect of oxygen deprivation accumulates through the biosynthesis chain. However, the significant changes in cholesterol biosynthesis could also be a result of promoting viral infectivity. Namely, host-derived lipids are required for the viral life cycle (i.e., entry, replication, and assembly) and could therefore also be a potential target for antiviral drug development [246, 248]. Higher level of plasma membrane cholesterol increases viral infection rate by promoting membrane fusion [248, 249], while its depletion disrupts virion membrane composition [250]. Additionally, Costello et al. [232] showed the importance of desmosterol for HCV replication by increasing membrane fluidity, while the inhibition of its biosynthesis resulted in an antiviral effect.

3.3.3 Computational models for estimating COVID-19 severity

Our machine learning models using a unique set consisting of only clinical information measured at hospital admission to predict COVID-19 disease course showed excellent performances (AUC: 0.96; Table 3.4; Figure 3.3A) comparable to other machine learning models published (Figure 3.3C) [251, 252]. Other biostatistical tools attempting to predict COVID-19 course have found sets of relevant variables and models comparable to our study. The neural network developed by Statsenko et al. [253] has demonstrated the ability to estimate the risk of patients being administered to the ICU, with ferritin, LDH, and CRP being powerful predictors. In an impressive and early work, Yan et al. [254] showed the stratification of patients at high and low risk of mortality by a tree-based model utilizing only CRP, LDH, and lymphocyte measurements. Xiong et al. [251] successfully trained several machine learning models to predict COVID-19 severity identifying a set of laboratory and imaging features as relevant. To put our model performances directly in context with established clinical practice, we compared them to the COVID-GRAM risk score [204], as well as a clinical baseline model we developed using three variables commonly used to stratify patients into disease severity groups (i.e., CRP, LDH, ferritin) (Figure 3.3C). It is evident that all of our models using clinical information significantly outperformed current clinical patient stratification strategies (AUC: 0.96 vs. $AUC_{baseline}$: 0.74 and $AUC_{COVID-GRAM}$: 0.68). Although the baseline model performed well despite only using three measurements, it is clear that by adding a few variables that are easy to assess and routinely measured, for example the reason for hospital admission or oxygen saturation levels, performances could be greatly improved. Unfortunately, the usefulness

of COVID-GRAM as a risk score predicting critical illness among patients hospitalized with COVID-19 could not be confirmed in our cohort. This finding supports the concerns initially raised by Moreno-Pèrez et al. [255] who found similar limitations in their Spanish cohort (AUC: 0.72). The inability to validate COVID-GRAM in European cohorts reflects the limitations encountered when applying risk prediction tools in new populations. In contrast to this, Sebastian et al. [256] have reported good correlation of COVID-GRAM derived risk scores with mortalities in a Polish cohort. This may indicate the usefulness of COVID-GRAM tool assessing the risk of fatality rate rather than disease severity.
A subset of 4 sterol measurements alone shows similar performance to COVID-GRAM risk estimates ($AUC_{T1sterols}$: 0.66; $AUC_{COVID-GRAM}$:0.68; Table 3.4; Figures 3.3A and 3.3C). However, when clinical variables were combined with sterol measurements, the performance of the model remained excellent without improving ($AUC_{clinical}$: 0.96; $AUC_{clinical+T1sterols}$: 0.95; Table 3.4; Figures 3.3A and 3.3C). We do not find it surprising since the clinical model already shows exceptional accuracies, reflecting a lack of performance improvement upon combining sterol measurements and clinical information. Although we do not find sterols to be indicative of disease severity, they should not be discarded as potentially useful biomarkers, as there are a number of other clinical outcomes highly relevant for optimal COVID-19 patient care. One of them gaining relevance with the rising number of long-COVID cases are the long-term adversarial effects of a SARS-CoV-2 infection, including on the liver. Independent of pre-existing chronic liver diseases, abnormalities in liver enzymes are common in COVID-19 patients, reflecting a dysregulation of hepatic function. In most patients without previous liver conditions hepatic injury is mild and transient [257], while a profound deterioration of hepatic injury was reported in patients with severe courses of COVID-19 [258]. Interestingly, studies also show elevation of the liver fibrosis index FIB-4 and serum hyaluronic acid in acute COVID-19 patients upon admission to hospital care, suggesting liver fibrogenesis [259]. Moreover, patients with chronic liver diseases and COVID-19 present with higher risk of long-term morbidity and fatality rate, where approximately 30% of those patients present with symptoms consistent with long COVID-19 [260]. Also, hospitalized patients had significantly lower HDL-cholesterol values at a follow-up [261]. A recent computational study shows that acute liver injury is a common complication in COVID-19 ($\tilde{3}$9.9%) with patients unable to fully recover until hospital discharge. The average time to recover may take up to two months, but can be reliably estimated using statistical models and measurements taken upon patients leaving the hospital [262]. All of this evidence suggests that the liver function, especially of metabolically compromised patients and those most susceptible for complications, should be monitored even after being discharged from hospital care.

3.4 Conclusion

In conclusion, infection-associated dyslipidemia in COVID-19 patients has been widely reported. However, most of the clinical studies concentrate only on lipoproteins [183, 184, 233] while little is known about the underlying mechanism of cholesterol metabolism. Herein we focused in depth on *de novo* intracellular biosynthesis of cholesterol in hospitalized COVID-19 patients and showed statistically significant differences in sterol concentrations over the course of COVID-19. A greater number of sterols were significantly altered in patients with a severe disease course (i.e., zymosterol, zymostenol 24-dehydrolathosterol, desmosterol, cholesterol), than in patients with a mild disease course (i.e., zymostenol, 24-dehydrolathosterol, cholesterol).

Furthermore, SARS-CoV-2 is known to be rapidly evolving, with symptoms and disease courses changing according to the most prevalent variant. The inability to validate risk scores in populations outside their original development, and the fact that much is still unknown about the long-term adverse effects on the health of recovered COVID-19 patients, warrants a continuing search for novel biomarkers characterizing this disease. Our machine learning models provided a unique set of 8 clinical variables sufficient to predict disease severity with excellent accuracy (AUC $= 0.96$). This proved to be a substantial improvement over currently used clinical risk scores. Although the concentrations of sterol intermediates have not improved our already exceptional clinical model, we have shown that their concentrations change during disease course and that the changes differ between mild and severe COVID-19 cases. Accordingly, we strongly believe that their biomarker potential should be further explored, as they may have prognostic value for clinical outcomes other than disease severity prediction, such as the adverse effects of a SARS-CoV-2 infection (including on the liver).

To our knowledge, this is the first study relating to COVID-19 to examine the blood sterol intermediates that arise from the endogenous biosynthesis of cholesterol in detail, which contributes to our understanding of the SARS-CoV-2 pathogenesis and disease course. The second contribution of our study is a unique set of readily available clinical variables capable of predicting COVID-19 course with excellent accuracy. Finally, we believe that our study will also serve as inspiration for future studies aimed at investigating potential biomarkers outside the routine clinical setting.

3.5 Limitations of the study

Because our study was based on hospitalized patients, the majority of the patients had severe disease. Thus, the distribution of patients in the present study reflected the actual situation in hospitalized patients but not in outpatients in whom mild(er) illness prevailed. In addition, since patients were admitted to hospital with different pre-hospital durations of illness and at different disease stages, baseline as well as consecutive blood samples were obtained over a considerable time span. Furthermore, the T2 samples were collected either at the occurrence of severe deterioration or in the middle of the hospitalization. Another limitation of our study is that some clinical variables could not be included because of the high proportion of missing data in patients. Furthermore, COVID-19 deterioration may not only be a result, of course, of SARS-CoV-2 infection, but is usually a consequence of a rather complex condition including underlying disease worsening, secondary infections, or noninfectious complications, that were not considered. Finally, it is essential to test our models in an external cohort and, although we found evidence of the potential value of sterols as biomarkers for COVID-19 and discussed their potential importance in the manuscript, we were not able to test these hypotheses in the context of this study.

3.6 Methods

3.6.1 Experimental model and study participant details

Human participants

165 adult patients (53 females and 112 males, aged 23 to 93 years) admitted to the Department of Infectious Diseases of the University Medical Center Ljubljana (Slovenia) from July 2020 to July 2021 with COVID-19, were enrolled in the prospective study. In all patients, infection with SARS-CoV-2 was demonstrated by the presence of the virus in nasopharyngeal swabs using real-time PCR. All participants provided written informed consent, and the study was approved by the Medical Ethics Committee of the Republic of Slovenia (No. 0120–211/2020/7 and No. 0120-33/2022/3).

3.6.2 Method details

Materials

All sterol standards (See Table S1 and key resources table) were bought from Avanti Polar Lipids (Alabaster, AL, USA), except cholesterol, which was purchased from Merck

(Darmstadt, Germany). LC-MS grade cyclohexane, methanol, and 1-propanol were purchased from Honeywell (Charlotte, NC, USA). Formic acid was from Fluka (Honeywell) and sodium hydroxide from Merck (Darmstadt, Germany).

Sample collection

Blood samples were collected at 3 different time points during the hospitalization of 62 COVID-19 patients: upon admission to hospital care due to a severe course of COVID-19, in case of severe deterioration or in the middle of treatment, and upon discharge from the hospital. In 103 COVID-19 patients, blood samples were only collected upon hospitalization. Samples were collected in vacutainer tubes and centrifuged at 1811 × g (Eppendorf 5810R) at room temperature for 10 min to obtain serum samples which were stored at -80°C until further use.

Assessment of clinical parameters

Information was obtained using a questionnaire. More than 200 clinical parameters were recorded, including patient demographics, comorbidities and associated diseases, regular therapies, COVID-19 vaccination status, symptoms and signs of the disease and their intensity, laboratory and X-ray findings, information regarding specific treatments, as well as a patient condition upon discharge. According to disease severity, patients were classified into 4 groups – mild, moderate, severe, and critical. Grouping was done according to NIH recommendations (Table S7) [263]. Demographic data and selected clinical and biochemical parameters are listed in Tables 3.1 and S2.

Sterol isolation

200 ng of lathosterol-D7 (Avanti Polar Lipids, Cat. No. 700056P) and 1 mL of hydrolysis solution (4g NaOH dissolved in 10 mL Milli-Q water and 90 mL 99.5% ethanol) were added to 250 µL of serum sample and mixed well. After 1 h of incubation in a water bath with shaking at 65°C, 500 µL of Milli-Q water and 3 mL of cyclohexane were added to the solution, vortexed, and centrifuged at 1301 × g (5810 R centrifuge, Eppendorf, Germany) for 10 min at room temperature. The upper organic phase was transferred to a new 15 mL glass vial and the extraction step with cyclohexane was repeated. Extracts were then combined and the organic solvent was evaporated at 45°C using Eppendorf concentrator 5301. Lipid films were dissolved in 100 µL of LC-MS grade methanol, transferred to HPLC vials, purged with N2, and stored at -20°C until LC-MS/MS analysis.

LC-MS/MS analysis

The analysis was carried out according to a modified method from our previous study [264]. Briefly, chromatographic separation using two pentafluorophenyl columns Phenomenex Luna 3 µm (Phenomenex, USA) was performed on a Shimadzu Nexera XR HPLC (Shimadzu, Japan), with an oven temperature of 40°C, mobile phase composition of methanol/1-propanol/formic acid/water (v/v/v/v, 80:10:0.05:9.95), and isocratic flow 200 µL/min, except for cholesterol 300 µL/min. The injected volume of standard or sample was 5 µL, except for cholesterol the injection volume was 1 µL.
Detection was performed on an SCIEX Triple Quad 3500 mass spectrometer (AB Sciex LLC, USA) with APCI ionization in a positive mode. Detailed information about sterols and mass spectrometry detection conditions are listed in Table S1. A standard solution consisting of the same concentration of each sterol was used for their quantification in serum samples. Analyst software 1.6.3 (AB Sciex LLC, USA) was used for data evaluation.

Data preprocessing

The selected target variable for predictions was the degree of disease severity, which was categorized into 4 classes, according to increasing severity and in consensus with NIH guidelines (Table S7) [263]. Class 1 represented the mildest and class 4 the most severe course of the disease. Due to the small number of members in classes 1, 2, and 4, patients were combined into two groups, namely those with mild illness (classes 1 and 2) and those with severe illness (classes 3 and 4) cases. Patients with missing disease severity annotations were discarded (N = 1), resulting in a sample of 30 and 134 patients for the mild and severe groups, respectively.
159 variables (also referred to as features throughout the manuscript) with the potential to be used for prediction were included in the information available upon hospital admission, namely patient demographics, vaccination status, comorbidities and associated diseases, regular therapies upon admission, symptoms and/or signs of the disease and their intensity, and other clinical and laboratory findings. Irrelevant or potentially biasing variables (e.g., dates, patient IDs, some clinical endpoints) were omitted during preprocessing. 77 variables were used as an input for variable selection.
Imputation was performed for variables if their degree of missingness did not exceed 15%, as imputation accuracy cannot be guaranteed in the case of higher missingness. Variables exceeding this threshold were removed from the analysis. During imputation, we accounted for the mixed data types present in the collected data by differentiating between continuous, binary, and categorical variables. Continuous features were treated

using the IterativeImputer from scikit-learn [265] and scaled through min-max normalization. Binary information was completed using the K-Nearest Neighbors method. Missing categorical features were imputed by the most frequently occurring value and encoded to numerical representation by an ordinal encoder.

Variable selection

To combat overfitting of the machine learning models, variables relevant for disease severity prediction were selected prior to model training using a variant of the unsupervised feature selection method proposed by Kursa et al. [266], called *Boruta*. Our variant of *Boruta*, referred as iterative *Boruta*, included the recording of relevant variables over 100 iterations of the feature selection process. Only variables occurring in at least 50% of iterations were deemed relevant and kept. This feature selection process was conducted separately for the clinical ("clinical") and sterol ("T1 sterols") datasets. Selected variables from both datasets were subsequently merged into a combined set ("clinical + T1 sterols").

Classification models

Eight predictive models were trained, namely Random Forest [267], Gaussian Processes [268], AdaBoost [269], Logistic Regression [270], K-Nearest Neighbors [271], Multilayer Perceptron [272], Gaussian Naive Bayes [273], and Quadratic Discriminant Analysis [274]. To ensure optimal performance, model hyperparameters were optimized during training through an exhaustive search, with balanced accuracy as a scoring metric.
The predictive power of each classification model was assessed by leave-one-out cross-validation (Figure S3) and evaluated using several metrics including balanced accuracy, precision, recall, F1-score, and ROC-AUC score. Their definition is based on a confusion matrix consisting of four elements (TP, true positive; TN, true negative; FP, false positive; FN, false negative). For additional information about each metric please refer to Note S1. Within each cross-validation split, the importance of individual variables for prediction was measured by applying an iterative permutation-based feature importance assessment [267] with 100 iterations. For analyses, the feature importance of all 100 iterations was averaged.

Comparison to COVID-GRAM and clinical baseline

COVID-GRAM, developed by Liang et al. [204], is a risk score for COVID-19 patients used to predict patients' risk of developing a critical illness. It utilizes 10 clinical variables available at the time of admission including age, the presence of hemoptysis, dyspnea, or abnormali-

ties in the X-ray, whether patients arrive unconscious, the number of comorbidities, records of cancer history, the neutrophil to lymphocyte ratio, and the concentrations of lactate dehydrogenase (LDH) and bilirubin. As the COVID-GRAM score was not assessed during patient recruitment, we calculated it post-hoc for each patient. Individual risk scores were subsequently translated into probabilities and used for further analyses. For respective formulas, please refer to the Note S2.

To make a comparison to current clinical practices, we selected three variables commonly used to stratify patients into disease severity groups, namely concentrations of LDH, ferritin, and C-reactive protein (CRP) upon admission. Classification models were subsequently trained using only these three variables (as outlined in section classification models) - we refer to those as “clinical baseline” models.

3.6.3 Quantification and statistical analysis

Statistical significance was tested comparing three time points of serum sterol intermediate concentrations with a nonparametric Friedman’s test since the collected data did not follow a normal distribution. For multiple comparisons, the adjusted p-values were determined using Dunn’s test. The adjusted p-value ≤ 0.05 was considered statistically significant. Statistical analysis was carried out using GraphPad Prism 9 software (Dotmatics, California, USA).

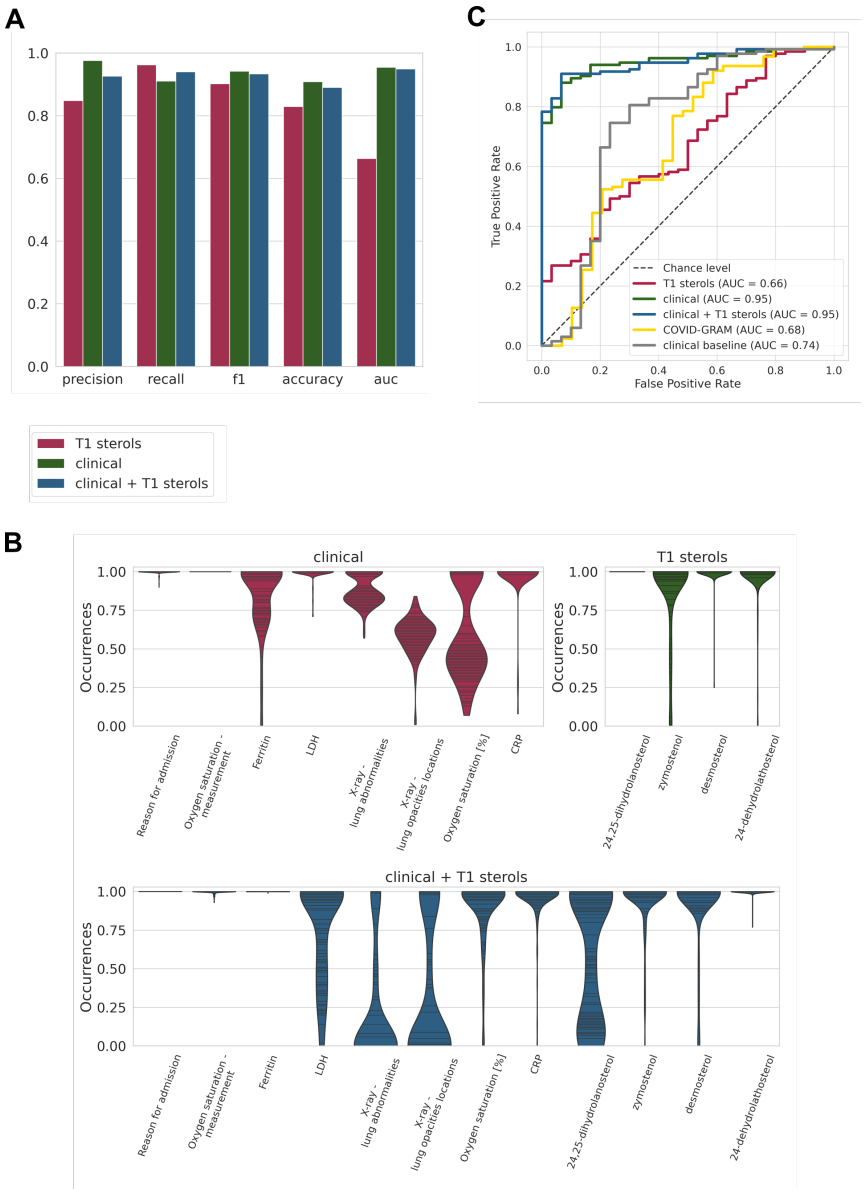


Figure 3.3: **Computational models for estimating COVID-19 severity**. (A) Performance evaluation of machine learning classification models for predicting disease severity in a cohort of COVID-19 patients (N = 164). This utilized three different sets of variables, namely clinical parameters (F = 8), sterol intermediates measured at T1 (F = 4), and their combination (F = 12). Only the best performing models for each variable set are displayed. F, number of features; N, number of patients. (B) Variable importance in different scenarios. Feature importance for all three scenarios is shown: upper left - clinical parameters; upper right - sterol intermediates in T1; lower: combination of both clinical and T1 sterols. ∗Oxygen saturation measured with or without oxygen supplementation upon admission. (C) Receiver operating characteristic (ROC) curve of all three scenarios, COVID-GRAM, and clinical baseline model. The area under the curve (AUC) for each set is displayed in the legend. The dashed diagonal line reflects the performance of a diagnostic test that is no better than chance level. See also Tables S5 and S6.

3.7 Supplementary Notes

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107799.

3.7.1 Data and code availability

Our code is available at https://github.com/sonjakatz/covid_sterols_ML. Data used for analysis in this study cannot be deposited in a public repository because of patient privacy concerns.

3.7.2 Acknowledgements

We would like to thank Tanja Blagus and prof. dr. Vita Dolžan (Pharmacogenetics Laboratory, Institute of Biochemistry and Molecular Genetics, Faculty of Medicine, University of Ljubljana), Petra Nassib, Jaka Brzin, and Jadranka Stojnić for the help with sample collection. Finally, we thank Benjamin Bajželj for critical input and carefully reading the manuscript, and John Hancock for appraisal of the manuscript.

3.7.3 Funding

This work was funded by the Slovenian Research and Innovation Agency (ARIS) program grants P1-0390, P2-0359, P3-0296, and the Ph.D. grant for young researchers to E.K.; S.K. was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860895 TranSYS. We would also like to acknowledge the support by the Network of Research and Infrastructure Centres of the University of Ljubljana (MRIC-UL-CFGBC, IP-0022), financed by the ARIS, and infrastructure program ELIXIR-SI RI-SE-2 financed by the European Regional Development Fund and by the Ministry of Education, Science and Sport of Republic of Slovenia. The funding sources had no role in the design of the study and collection, analysis, and interpretation of data nor in writing the manuscript.

Chapter 4

A validated model for early prediction of group A streptococcal aetiology and clinical endpoints in necrotising soft tissue infections

Sonja Katz, Jaco Suijker, Steinar Skrede, Annebeth Meij-de Vries, Anouk Pijpe, Anna Norrby-Teglund, Laura M Palma Medina, Jan K Damås, Ole Hyldegaard, Erik Solligård, Mattias Svensson, PerAID/PerMIT/INFECT study group, Knut Anders Mosevoll, Vitor AP Martins dos Santos, and Edoardo Saccenti

Manuscript submitted

Abstract

Objectives: To develop and externally validate machine learning models for predicting microbial aetiology and clinical endpoints, encompassing surgery, patient management, and organ support in Necrotising Soft Tissue Infections (NSTI).

Methods: Predictive models for the presence of Group A Streptococcus (GAS) and for five clinical endpoints (risk of amputation, size of skin defect, maximum skin defect size, length of ICU stay, and need for renal replacement therapy) were built and trained using data from the prospective, international INFECT cohort (409 patients, 2013-2017) implementing unsupervised variable selection and comparing several algorithms. SHapley Additive exPlanations (SHAP) analysis was used for model interpretation. GAS predictive models were externally validated using data from a Dutch retrospective multi-centre cohort from the same calendar period (216 patients).

Results: Eight variables available pre-surgery (age, diabetes, affected anatomical locations, prior surgical interventions, and creatinine and haemoglobin levels) sufficed for prediction of GAS aetiology with high discriminatory power in both the development (ROC-AUC: 0.828; 95%CI 0.763, 0.883) and validation cohort (ROC-AUC: 0.758; 95%CI 0.696, 0.821). The prediction of clinical endpoints related to surgical, patient management, and organs support aspects, was not successful.

Conclusion: An externally validated prediction model for GAS aetiology before organ support aspects was unsuccessful, having implications for targeted treatment decisions of NSTI.

4.1 Introduction

Necrotising soft tissue infections (NSTI) are rare, yet severe, bacterial infections that impact a diverse demographic population regarding age, gender, and comorbidities [127]. NSTI can be caused by either a single pathogen (monomicrobial, type II NSTI), most often *Streptococcus pyogenes* (group A streptococcus; GAS), or a combination of multiple pathogens (polymicrobial, type I NSTI), each employing distinct pathogenic mechanisms [126, 128]. In addition to tissue damage, NSTI is characterised by systemic toxicity, often resulting in sepsis and septic shock (28-50%) [19, 130]. The course of the disease is rapid, and mortality rates vary from 10-29% [130, 132, 134]. Long-term consequences include extensive morbidity, manifesting in functional impairments such as amputations, scarring and fatigue with significant psychosocial impacts, including fear of recurrence, post-traumatic stress, depression, and changes in social engagement [137, 139].

Given the severity of NSTI, an early overview of disease characteristics predicting outcomes could significantly improve prognosis and outcomes. A major factor in deciding the appropriate treatment revolves around ascertaining whether GAS is the causative microorganism [275], thereby guiding specific antimicrobial treatment including clindamycin [276]. Whereas intravenous immunoglobulin (IVIG) is not routinely recommended for the treatment of NSTI in general, data supports it may lead to improved outcomes in case of GAS based NSTI [277, 278]. Therefore some guidelines recommend routinely using [279] or considering IVIG [275] in case of GAS-based NSTI. Because the earliest method to identify the likelihood of GAS involvement is a gram stain on samples obtained during surgery [279], an earlier accurate, and preferably pre-surgery, identification of these patients could reduce the time until the start of IVIG within hours. Given the recent worldwide surge in the incidence of GAS [280–282], an early prediction of GAS aetiology could lead to the start of targeted treatment before results of microbiological tests are available, while avoiding over-treatment.

Applying machine learning-based predictive models holds promising potential in overcoming shortcomings related to accuracy, individualisation, and applicability in clinical decision-making for NSTI patients. We have recently shown (18) that such a model can accurately predict 30-day mortality. We built and trained a predictive machine learning model using just sixteen clinical parameters/variables collected within the initial 24 hours of ICU admission enabling a more precise prediction of 30-day mortality compared to commonly used clinical scoring systems, like the Simplified Acute Physiology Score (SAPS) [148] and the Sequential Organ Failure Assessment (SOFA) [146] scores used to predict mortality in intensive care units (ICU) patients.

In this study we develop and externally validate prediction models for several clinical endpoints relevant to NSTI, including likely GAS aetiology, surgical aspects (risk of amputation, size of skin defect after fist surgery, maximum skin defect size), as well as patient management (length of ICU stay), and need for organ support (need for renal replacement therapy) .

4.2 Methods

4.2.1 Study design

Development cohort

Subjects and data of the cohort used to develop and train the prediction models were obtained from the INFECT study (https://permedinfect.com/; registration number NCT01790698, ClinicalTrials.gov), an international, multi-centre, prospective, cohort study with patients with NSTI included prospectively at five Scandinavian hospitals (Supplementary Note 1). A total of 409 patients above the age of 18 and with surgically confirmed NSTI cases were enrolled between February 2013 and June 2017. Patients' enrolment and data collection protocols, as well as demographics, have been published previously [19, 126].

Validation cohort

Subjects and the predictive models were obtained from a Dutch retrospective multi-centre cohort [283] comprised of 216 patients admitted for acute treatment of NSTI to 11 centres between January 1, 2013, and December 31, 2017, irrespective of a subsequent ICU admissions (Supplementary Note 1). Patient characteristics have been previously described [283].

4.2.2 Clinical endpoints

Six clinically relevant NSTI endpoints were selected through semi-structured interviews as previously described [284] and were used for machine learning modelling. The selected endpoints encompass different clinical aspects of NSTI. Bacterial aetiology: presence of GAS (binary). Surgical aspects: risk of amputation (binary); size of skin defect after fist surgery, % of body surface (continuous); maximum skin defect size, % of body surface (continuous). Patient management: length of ICU stay (continuous). Organ support: need for Renal Replacement Therapy (RRT) (binary). The binary endpoints were coded as 0-1

Table 4.1: **Overview for the number of patients included and data subsets assessed for each clinical outcome.**

Outcome category	Outcome	No. of patients (n)	Time-dependent data subsets	Prediction task	No. of patients (%) or mean ± SD
causative microbes	Presence of GAS	409	Entry, pre-surgery, post-surgery, baseline	classification (y/n)	126 (30.8%)
surgical aspects	Risk of amputation	409	Entry, pre-surgery	classification (y/n)	54 (13.2%)
	Size of skin defect (after first surgery)	391	Entry, pre-surgery	regression (pct. of body surface)	4.9 ± 5.2
	Size of skin defect (maximal)	409	Post-surgery, baseline	regression (pct. of body surface)	6.4 ± 7.3
patient management	Length of ICU stay	402	Entry, pre-surgery, post-surgery, baseline	regression (days)	10.7 ± 10.8
organ support	Need for RRT (within 24h after ICU admission)	409	Entry, pre-surgery, post-surgery	classification (y/n)	57 (13.9%)
	Need for RRT (within 90 days)	351	Baseline	classification (y/n)	24 (6.8%)

(absence-presence, no-yes). An overview on the number of patients included and label balance for each clinical outcome can be found in Supplementary Table 4.1.

4.2.3 Data preprocessing

For our development cohort we utilised data from within the INFECT study during which over 700 clinical variables were collected, encompassing data available from hospital admission to ICU admission [126]. The clinical variables were categorised depending on their availability at four time points: *entry* (upon hospital admission, 45 variables), *pre-surgery* (prior to the first surgical procedure in the referral centre, 56 variables), *post-surgery* (posterior to first surgical procedure and prior to ICU admission, 723 variables), *baseline* (BL; first 24 h of ICU admission, 762 variables). For a graphical overview detailing the time point for data collection see Supplementary Figure 4.1b; the number of variables in each subset can be found in Supplementary Table 2. Data cleaning, imputation, and post-processing have been implemented as previously described [284]. For validation of models, we extracted the subset of variables required for prediction from the validation cohort and preprocessed the data using the same cleaning and imputation procedure as those applied to the development cohort.

4.2.4 Selection of input variables for the prediction

The selection of variables relevant for the prediction of each clinical endpoint was done through unsupervised variable selection utilising the Boruta algorithm [164]. To obtain robust sets of relevant variables, we implemented an iterative version. Therein, we i) created a subset of the dataset by randomly sampling 80% of patients and ii) iteratively ran Boruta for 50 times with different initialisation on the bootstrapped dataset and iii) repeated steps i) - ii) for 30 iterations. Only variables identified in more than 50% of the

iterations were considered relevant and kept for further analyses.

4.2.5 Machine-learning model development

A graphical overview detailing the developmental and validation process of the machine learning models can be found in Supplementary Figure 4.1a.

Model selection

For classification models involving binary endpoints (GAS aetiology, need for RRT, risk of amputation), three different predictive models based on different statistical and machine learning approaches were developed and compared: logistic regression [285], Gaussian process classifier [286], and Random Forest classifier [74]. Models for regression tasks (length of ICU stay, maximum size of skin defect) included lasso regression, ridge regression, elastic nets, Gaussian process regression [286], multi-layer perceptron regression, and Random Forest regression [74].
The predictive power of each model was compared through leave-one-out cross-validation. Model hyperparameters were optimised using an exhaustive grid search, using balanced accuracy for classification and negative mean squared error for regression as scoring metric for model selection; a summary on the hyperparameters optimised can be found in Supplementary Note 2. Only the best performing models were used for internal validation with bootstrapping.

Model training and internal validation

Internal validation was conducted to ensure optimism-corrected performance evaluation and quantify the uncertainty associated with our developed models. This entailed repeating the model training and hyperparameter optimisation process over 1000 bootstrap samples (n=1000) drawn from the development cohort, enabling the calculation of robust 95% confidence intervals (95% CIs).

External validation

For external validation, the generalisation performance of the trained models was assessed through bootstrapping of the validation cohort (n = 10000). To enable models to account for the heterogeneity of data across time, geography and facilities, we adopted the local validation scheme proposed by [287] and fine-tuned models trained on the development cohort using 30% of the validation data.

Model performance

The quality of The performance of classification models was evaluated using several metrics: Precision, Recall, F_1-score, balanced accuracy, Brier score, area under the receiver-operator curve (ROC-AUC), and average precision-recall score (PR). The quality of regression models was assessed using the coefficient of determination (R^2) as well as the mean absolute error (MAE) and mean squared error (MSE). Quality measures are given as mean and associated 95%CI calculated over all bootstrapping iterations.

Model explainability

SHapley Additive exPlanations (SHAP) values were used to assess relative contribution of clinical variables to the prediction/classification models [288]. SHAP values were calculated for each model training/validation step since they are sensitive to model parameterisation and data splits. Results are given as mean with associated 95%CI.

Software

For all machine-learning models the implementations available in the *scikit-learn* Python library (version 1.4.2) were used. For variable selection, Boruta version 0.3 was used [164]. SHAP analysis was conducted using the package version 0.43.0.

4.3 Results

4.3.1 Prediction of GAS aetiology in NSTI

To predict GAS aetiology in NSTI as early as possible, we trained prediction models for the presence/absence of GAS using time-dependent subsets of clinical variables, namely *entry*, *pre-surgery*, *post-surgery*, *baseline*. We conducted variable selection to identify a minimal set of clinically relevant variables that are feasible to obtain in a clinical setting. This process aimed to optimise the model by reducing the risk of overfitting while maintaining high prediction quality. The number of selected variables ranged from 6 (*entry*) to 14 (*baseline*) as shown in Table 4.2.
Based on these reduced sets of variables, the performance of several machine learning-based classifiers was compared, with Random Forest classifiers standing out in terms of performance across all subsets. Therefore, only results for the Random Forest prediction model will be presented. Results for other models are available in Supplementary Table 4.

Table 4.2: **Overview of variables yielded through unsupervised variable selection**. A more detailed variable description can be found in Supplementary Table 3.

Entry (6/45 variables)	**Pre-surgery** (8/56 variables)	**Post-surgery** (9/723 variables)	**Baseline** (14/762 variables)
affected[1]: upper arm	affected[1]: upper arm	affected[1]: upper arm	affected[1]: upper arm
affected[1]: lower arm	affected[1]: lower arm	affected[1]: lower arm	affected[1]: lower arm
affected[1]: anogenital area	affected[1]: anogenital area	affected[1]: anogenital area	affected[1]: anogenital area
surgery before	surgery before	surgery before	surgery before
diabetes	diabetes	diabetes	diabetes
age	creatinine[2]	creatinine[2]	creatinine[2]
	haemoglobin[2]	haemoglobin[2]	haemoglobin[2]
	age	creatinine[3]	creatinine[3]
		Anatomical site sampled	systolic BP (lowest)[4]
			creatinine[4]
			noradrenaline[4]
			platelets[4]
			lactate[4]
			glucose[4]

[1] at arrival at specialised hospital

[2] preoperative (preop): before the first surgery, which is before ICU admission

[3] preadmission: upon ICU admission

[4] baseline (BL): during the first 24 hours in the ICU

Comparison of prediction performances showed distinct differences between different time-dependent subsets (Supplementary Table 4.3). Notably, this revealed that performances peak using information already available *pre-surgery* (ROC-AUC 0.828; 95%CI 0.763, 0.883), constituting the earliest possible time point for prediction of GAS aetiology (Figure 4.2a, blue).

We sought to validate the performance of the *pre-surgery* model in an external validation cohort, composed of 208 patients with information available on microbial aetiology (208/216 patients, 96.3%). A comparison of the variable characteristics between the development and validation cohorts showed a high degree of similarity (Table 4.4).

Overall, our model showed a good discriminatory power within the validation cohort with a ROC-AUC of 0.727 (95%CI 0.672, 0.783) (Figure 4.2a, green). Fine-tuning trained models using 30% of the validation data improved its performance to a ROC-AUC of 0.758 (95%CI 0.696, 0.821) (Figure 4.2a, pink; Supplementary Table 4.3). Determination of the optimal threshold by trying to minimise the number of false negatives, yielded an ideal cutoff value of 40%, resulting in a good trade-off between sensitivity and specificity (Table 4.5).

Table 4.3: **Model performances in estimating GAS involvement for time-dissected datasets**. Depicted are the mean and the 95% confidence intervals (in parentheses). Acc.: balanced accuracy, Prec.: precision, Brier: Brier score, ROC AUC: area under the receiver-operator curve, Ave. prec.: average precision

	Entry	Pre-surgery	Post-surgery	Baseline	Pre-surgery (fine-tuned)
Acc.	0.652 (0.574,0.735)	0.726 (0.642,0.796)	0.723 (0.644,0.791)	0.727 (0.658,0.799)	0.677 (0.617, 0.737)
Prec.	0.572 (0.435,0.724)	0.666 (0.548,0.784)	0.683 (0.556,0.811)	0.729 (0.600,0.867)	0.644 (0.534, 0.776)
Recall	0.463 (0.267,0.698)	0.583 (0.391,0.739)	0.564 (0.396,0.720)	0.547 (0.396,0.700)	0.568 (0.417, 0.726)
F1-score	0.502 (0.351,0.632)	0.617 (0.483,0.719)	0.613 (0.486,0.714)	0.621 (0.500,0.729)	0.598 (0.509, 0.682)
Brier	0.170 (0.142,0.203)	0.152 (0.128,0.183)	0.149 (0.125,0.180)	0.147 (0.125,0.172)	0.199 (0.168, 0.238)
ROC AUC	0.794 (0.733,0.851)	0.828 (0.763,0.883)	0.836 (0.775,0.891)	0.839 (0.779,0.894)	0.758 (0.697, 0.817)
Ave. prec.	0.617 (0.494,0.719)	0.684 (0.568,0.787)	0.685 (0.570,0.788)	0.703 (0.592,0.807)	0.701 (0.604, 0.780)

Table 4.4: **Comparison of predictive variables' characteristics between the development and validation cohort**. Only variables relevant for the *pre-surgery* model were considered. Detailed variable descriptions can be found in Supplementary Table 3.

	Development cohort (n=409)	Validation cohort (n=208)
GAS, n (%)	126 (30.8)	82 (39.4)
age, y, mean (SD)	58.7 (15.1)	58.2 (15.5)
affected: upper arm, n (%)	48 (11.7)	20 (9.3)
affected: lower arm, n (%)	56 (13.7)	23 (10.7)
affected: anogenital area, n (%)	143 (35.0)	64 (29.8)
surgery before, n (%)	77 (18.8)	53 (26.2)
diabetes, n (%)	98 (24.0)	58 (26.9)
creatinine (preop), mean (SD)	161.2 (121.3)	153.7 (123.3)
haemoglobin (preop), mean (SD)	11.4 (3.04)	12.7 (2.5)

4.3.2 Model interpretation

To gain more insight into the decision-making of the model and quantify how much individual variables contributed to predictions, we conducted post-hoc interpretability analysis using SHapley Additive exPlanations (SHAP) (Figure 4.2b). Inspecting the SHAP values revealed the profound impact of anatomical location on model decisions, with infections occurring in upper extremities being highly predictive of GAS, whereas infections in the anogenital areas hinted at a non-GAS infection. The preoperative creatinine levels, however, appears to be the most influential variable, with values above approximately 110 µmol/L being indicative of GAS (Supplementary Figure 3a). On the other hand, the presence of diabetes, a surgery conducted four weeks prior to diagnosis, and a higher age (>50

Threshold [%]	Sensitivity (TPR)	Specificity (TNR)	FPR	FNR
0	1.00 (1.00, 1.00)	0.00 (0.00, 0.00)	1.00 (1.00, 1.00)	0.00 (0.00, 0.00)
10	0.93 (0.81, 1.00)	0.29 (0.11, 0.49)	0.71 (0.51, 0.89)	0.07 (0.00, 0.19)
20	0.85 (0.69, 0.97)	0.46 (0.29, 0.63)	0.54 (0.37, 0.71)	0.15 (0.03, 0.31)
30	0.76 (0.59, 0.90)	0.59 (0.42, 0.74)	0.41 (0.26, 0.58)	0.24 (0.10, 0.41)
40	**0.67 (0.50, 0.82)**	**0.69 (0.55, 0.84)**	**0.31 (0.16, 0.45)**	**0.33 (0.18, 0.50)**
50	0.57 (0.42, 0.73)	0.79 (0.64, 0.91)	0.21 (0.09, 0.36)	0.43 (0.27, 0.58)
60	0.46 (0.31, 0.62)	0.86 (0.73, 0.97)	0.14 (0.03, 0.27)	0.54 (0.38, 0.69)
70	0.36 (0.21, 0.52)	0.92 (0.80, 0.99)	0.08 (0.01, 0.20)	0.64 (0.48, 0.79)
80	0.25 (0.10, 0.41)	0.96 (0.87, 1.00)	0.04 (0.00, 0.13)	0.75 (0.59, 0.90)
90	0.13 (0.02, 0.29)	0.99 (0.95, 1.00)	0.01 (0.00, 0.05)	0.87 (0.71, 0.98)
100	0.00 (0.00, 0.03)	1.00 (1.00, 1.00)	0.00 (0.00, 0.00)	1.00 (0.97, 1.00)

Table 4.5: **Summary of fine-tuned model accuracy at different decision threshold levels**. Depicted are the mean and the 95% confidence intervals (in parentheses). Highlighted in bold is the threshold identified as optimal trade-off between model precision and recall when aiming to reduce false-negatives as much as possible. The FPR can be interpreted as the risk of over-treatment, while the FNR indicates the risk for under-treatment. TPR - true positive rate, TNR - true negative rate, FPR - false positive rate, FNR - false negative rate.

years; Supplementary Figure 3b) were pointed at a low risk for GAS involvement. SHAP findings were consistent between development and validation cohort (Supplementary Figure 4).

4.3.3 Predicting clinical endpoints: surgical aspects, patient management, and need of organ support

To explore the possibility of estimating clinical endpoints beside the causative microorganisms, we trained predictive models to estimate surgical aspects (risk of amputation, the size of skin defect after fist surgery, the maximal size of skin defect), patient management aspects (such as the length of ICU stay), as well the necessity of organ support (need for RRT within 24 hours after ICU (BL)).

For each outcome, we conducted unsupervised variable selection, yielding between 11 and 16 predictive variables (Supplementary Table 6). Similar to the prediction of GAS aetiology, the performance of several machine-learning models on all time-dependent subsets were compared. Internal validation of the best performing models revealed a lack of predictive performance across all clinical endpoints (Figure 4.3a, b). Analysis of SHAP values for surgery-related endpoints, such as the risk of amputation, revealed that models assign high importance to clinically relevant variables, such as the presence of discoloration, creatinine

Table 4.6: **Prediction performance for clinical endpoints revolving around surgical aspects, patient management, and organ support.** Ave. prec. : average precision, R^2 coefficient of determination, MAE: mean absolute error, MSE: mean squared error

	Risk of amputation	Size of skin defect (after first surgery)	Size of skin defect (maximal)	Days spent in ICU	Need for RRT (24h after ICU admission)
Acc.	0.503 (0.483, 0.546)	-	-	-	0.543 (0.493, 0.610)
Prec.	0.137 (0.000, 1.000)	-	-	-	0.467 (0.000, 1.000)
Recall	0.020 (0.000, 0.106)	-	-	-	0.109 (0.000, 0.250)
F1-score	0.033 (0.000, 0.182)	-	-	-	0.168 (0.000, 0.345)
Brier	0.113 (0.092, 0.136)	-	-	-	0.105 (0.083, 0.125)
Ave. prec.	0.261 (0.159, 0.411)	-	-	-	0.385 (0.238, 0.540)
R^2	-	0.120 (0.044, 0.178)	0.179 (0.100,0.234)	0.018 (-0.063,0.056)	-
MAE	-	3.500 (3.144, 3.831)	4.383 (3.905, 4.879)	7.047 (6.194,7.976)	-
MSE	-	24.086 (16.334, 34.611)	43.328 (27.789, 66.054)	111.952 (65.077, 172.308)	-

values, of patient's age (Figure 4.3c).

4.4 Discussion

In this study we developed and externally validated prediction models for aetiology and various clinical endpoints related to NSTI. Our findings indicate that by using variables available before the first surgery in the referral centre, the presence of GAS aetiology can be successfully predicted, which was confirmed during external validation. Additionally, we showed that fine-tuning trained models on a fraction of the validation data further improved model performances without leading to over-fitting. However, other clinical endpoints explored in this study (i.e. amputation of an extremity, size of the skin defect after surgery, ICU length of stay, need for Renal Replacement Therapy (RRT)) could not be satisfactorily predicted.

The eight variables relevant to predicting GAS selected through unsupervised variable selection proved to be diverse, ranging from the anatomical location involved to laboratory values. However, all of the selected variables have previously been associated with GAS infections. Thus, prior investigations in NSTI have revealed a higher occurrence of monomicrobial infections, including GAS, in the upper extremities, while polymicrobial infections were more common in the truncal region [289]. This study found a negative association between other surgeries performed last four weeks prior to the NSTI and GAS etiology. There were 77/409 (19%) patients in the INFECT study with this risk factor [19], whereas numbers were 5/126 (4%) in patients with GAS etiology [277]. The elevated creatinine among GAS NSTI cases reflects that septic shock (65%) and multiorgan dysfunction are common in this group and more frequent than in NSTI of other microbial etiologies [277]. Differently, among the GAS NSTI cases there is a negative association to hemoglobin concentrations. This may be explained by the higher number of patients

with pre-existing comorbidities, more lengthy clinical courses, and more surgery in the preceding four weeks in the non-GAS cases [19, 277].

Our assessment of surgical endpoints, such as the risk for amputation and the size of the skin defect, highlights the challenge in objectively evaluating surgical endpoints. Interestingly, we noted that the variables deemed important in predicting the risk of amputation, such as discoloration, bruising, or age, appear to be clinically closely linked to surgical endpoints [161]. Despite the logical relevance of these variables, they prove insufficient in predicting skin defect size and whether an amputation of an extremity is performed. This suggests that either unsupervised variable selection failed to yield objective predictors, and more detailed information about surgical findings upon first debridement would be required. Alternatively, the variability in the decision to perform amputation and the amount of skin that needs to be excised are subjective and influenced by individual surgeons or local guidelines. This latter argumentation is supported by a recent interactive survey on current practice of the debridement of NSTI, which found a major variety in the amount of skin that needed to be resected according to Dutch general surgeons and plastic surgeons [290]. These results may imply a similar variability in the decision-making process surrounding the need for amputation. A gene predisposition in the STING gene has been linked to amputation and associated with the expression of virulence factors, indicating the complex interaction between host and pathogen influencing NSTI patients' outcomes [291].

Solely using clinical information ranging from hospital admission to the first 24 hours in ICU, we were unable to accurately estimate the length of stay in the ICU. Given the complexity and heterogeneity of NSTI progression and the fact that patients may need to spend prolonged periods in ICU [19], we believe more longitudinal ICU data is required to predict this outcome successfully.

Previous studies have proven the value of providing decision-support regarding patients' renal status [170] in predicting the need for renal replacement therapy (RRT). However, the limited number of patients requiring RRT in INFECT resulted in wide confidence intervals, which hindered our ability to draw firm conclusions on the efficacy of our models.

The combination of clinical and computational biology expertise presents a significant strength of this study. By employing a data-driven approach under the guidance of clinical experts, we were able to investigate previously unexplored clinically significant endpoints for NSTI. In addition to utilising the largest NSTI cohort currently available, we conducted rigorous internal validation, as well as externally validated our findings using data from a referral centre not included in the original study. This allowed us to realistically estimate the generalisability of our models. However, the usefulness of external validation to measure a model's clinical utility and generalisability across different institutions has

recently been questioned [287, 292]. By fine-tuning a pre-trained model with local data, we fully exploited the transferability of machine learning-based models. Our approach showcased the significance of conducting local fine-tuning, enabling the models to adapt to the potential heterogeneity present in data across different timeframes, geographical locations, and facilities, all while avoiding overfitting. The strengths of our findings should be considered within the context of certain limitations. The development cohort utilized originates from an ICU-focused study, with limited access to pre-hospital data. Due to the uncertainty surrounding the timing of initial symptoms, there exists the possibility of considerable diversity in the progression of the disease among patients, which could potentially affect the performance of our models. Also, we believe the inability to estimate patient management and need of organ support can be partially attributed to the lack of longitudinal data and large imbalances in target labels. Lastly, despite successful external validation, the clinical usefulness of models must still be assessed through prospective validation directly comparing model results with clinician's decision.

Our findings demonstrate that prediction of bacterial aetiology is possible, which opens up the possibility of delivering targeted interventions earlier in the disease course of NSTI. The early availability of predictive variables implies that our models can already be used in an emergency room setting. Given the rising incidence of GAS in the Western world [280–282] and studies indicating the beneficial effects of early IVIG administration on survival in GAS patients [277], we believe our findings hold high clinical relevance. Additionally, the conclusions drawn from the prediction of clinical endpoints highlight the need for more research on surgical decision-making.

To the best of our knowledge, this is the first study using machine learning-based methods to estimate aetiology and clinical endpoints relevant to improving NSTI care. Using only eight readily available variables, we developed and validated models capable of estimating the bacterial aetiology prior to surgical debridement. We believe the results of this study to have significant implications for sepsis treatment in patients with NSTI caused by GAS, which may ultimately improve their survival and quality of life.

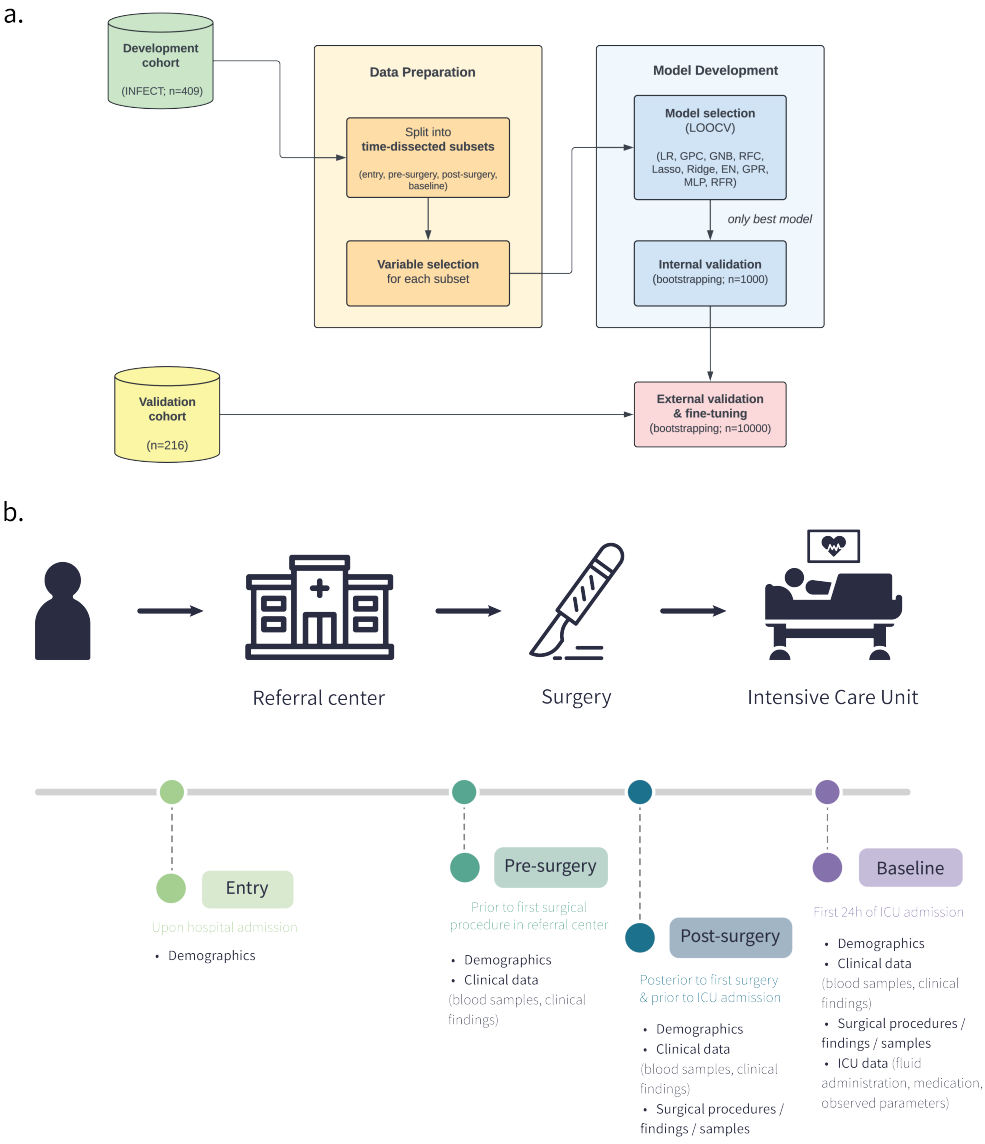


Figure 4.1: **Structured flowcharts detailing the (a) developmental and validation process of the machine learning models (b) time-dependent data dissection.** (a) This conceptual pipeline was systematically implemented for every clinical outcome assessed. External validation was exclusively pursued concerning the bacterial etiology (presence of Group A Streptococcus).(b) The INFECT study cohort was split into the time-dependent subsets entry, pre- and post-surgery at referral center, baseline, depending on the clinical availability of the data. n: number of patients/iterations; LOOCV: leave-one-out cross-validation; LR: logistic regression, GPC: Gaussian process classifier, GNB: Gaussian naive bayes; RFC: Random Forest classifier; Lasso: Lasso regression, Ridge: Ridge regression, EN: elastic net regression, GPR: Gaussian process regression, MLP: multi-layer perceptron; RFR: Random Forest regression.

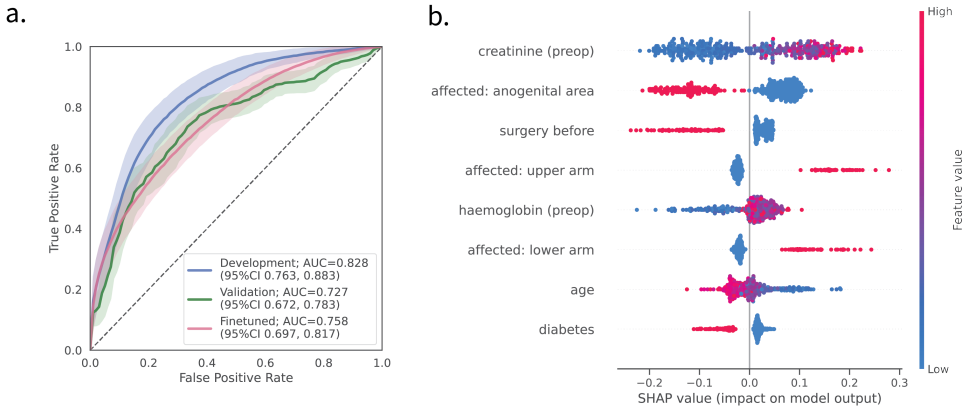


Figure 4.2: (a) **ROC-curves** comparing the performance of the development (blue), external validation (green), and fine-tuned validation (pink) cohorts. 95%CI: 95% confidence interval derived through 1000 (development) and 10000 (validation) bootstrapped samples (b) **SHAP values for models estimating of GAS aetiology**. Variables are sorted from most impactful (top, creatinine) to least impactful (bottom, diabetes), with every dot representing a patient. Positive SHAP values for a variable indicate a positive contribution to the model's decision to identify the patient as GAS-positive. Conversely, negative SHAP values indicate a contribution to classifying the patient as GAS-negative. The colour gradient denotes the variable values, with red indicating high values (e.g. age 70 years) and blue indicating low values (e.g. age 20 years). SHAP findings are consistent between development and validation cohort.

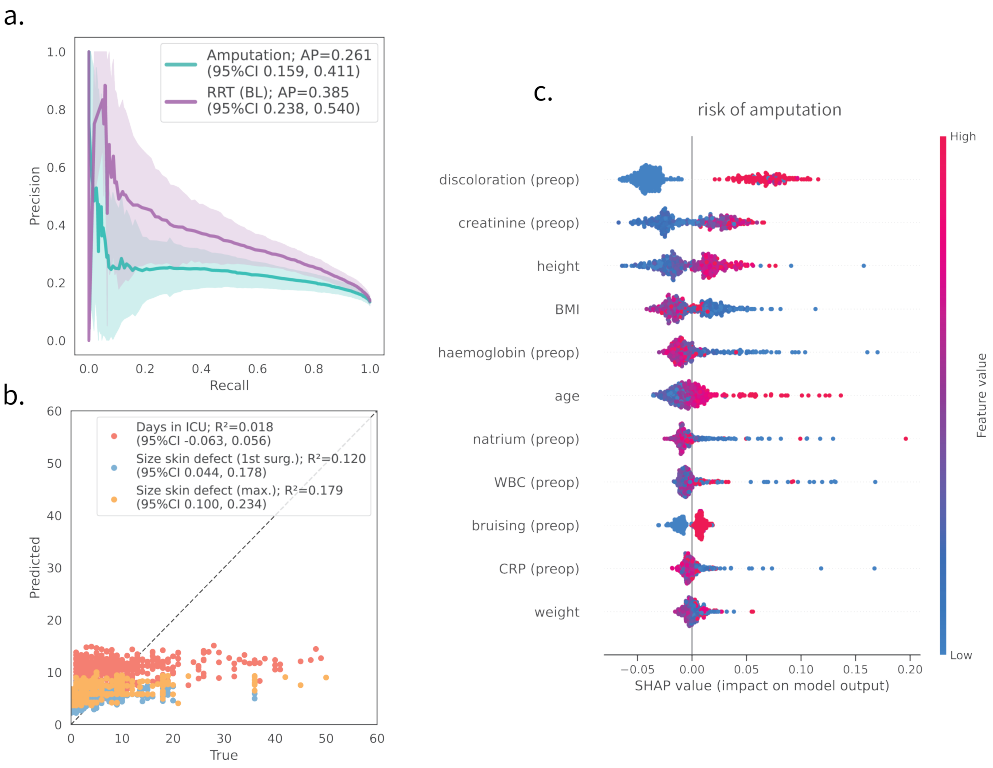


Figure 4.3: **Prediction performance for clinical endpoints revolving around surgical aspects, patient management, and organ support.** (a) Precision-recall curves for binary endpoints, including predicting the risk of amputation (turquoise) and need for RRT within the first 24h after ICU admission (baseline, BL). (b) Predicted versus true values for continuous endpoints, including the estimated days in ICU (red, days), size of skin defect after first surgery (blue, percent body surface), and the maximal size of skin defects (orange, percent body surface). (c) SHAP values for models predicting the risk of amputation. Depicted are the mean and the 95% confidence intervals derived through 1000 bootstrapped samples (in parentheses). Summary on included variables and more performance metrics can be found in Table 4.6 and Supplementary Table 6 respectively. AP: average precision, R^2 coefficient of determination.

4.5 Supplementary Notes

Supplemental information can be found online at `10.5281/zenodo.11517539`.

4.5.1 Data and code availability

Our source code for the GAS predictive models is available at `https://github.com/sonjakatz/permit-nsti-gas`.

4.5.2 Acknowledgements

The authors are grateful to the members of the INFECT, PerAID and PerMIT projects, for fruitful discussion on the manuscript.

4.5.3 Funding

This study has received funding from 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, 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 personalised medicine—smart combination of pre-clinical and clinical research with data and ICT solutions; and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860895 TranSYS.

Part 2

Patient Stratification

Chapter 5

Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping

Zuqi Li[†], **Sonja Katz**[†], Edoardo Saccenti, David W. Fardo, Peter Claes, Vitor A.P. Martins dos Santos, Kristel Van Steen, and Gennady V. Roshchupkin

[†]authors contributed equally

Published in: bioRxiv. 2024 Feb 8:2024-02.
DOI: 10.1101/2024.02.05.578873

Abstract

Unsupervised learning, particularly clustering, plays a pivotal role in disease subtyping and patient stratification, especially with the abundance of large-scale multi-omics data. Deep learning models, such as variational autoencoders (VAEs), can enhance clustering algorithms by leveraging inter-individual heterogeneity. However, the impact of confounders - external factors unrelated to the condition, e.g. batch effect or age - on clustering is often overlooked, introducing bias and spurious biological conclusions. In this work, we introduce four novel VAE-based deconfounding frameworks tailored for clustering multi-omics data. These frameworks effectively mitigate confounding effects while preserving genuine biological patterns. The deconfounding strategies employed include: i) removal of latent features correlated with confounders ii) a conditional variational autoencoder, iii) adversarial training, and iv) adding a regularization term to the loss function. Using real-life multi-omics data from TCGA, we simulated various confounding effects (linear, non-linear, categorical, mixed) and assessed model performance across 50 repetitions based on reconstruction error, clustering stability, and deconfounding efficacy. Our results demonstrate that our novel models, particularly the conditional multi-omics VAE (cXVAE), successfully handle simulated confounding effects and recover biologically-driven clustering structures. cXVAE accurately identifies patient labels and unveils meaningful pathological associations among cancer types, validating deconfounded representations. Furthermore, our study suggests that some of the proposed strategies, such as adversarial training, prove insufficient in confounder removal. In summary, our study contributes by proposing innovative frameworks for simultaneous multi-omics data integration, dimensionality reduction, and deconfounding in clustering. Benchmarking on open-access data offers guidance to end-users, facilitating meaningful patient stratification for optimized precision medicine.

5.1 Introduction

Unsupervised learning, in particular clustering, focuses on subgrouping individuals based on their intrinsic data structures, therefore playing an essential role in tasks like disease subtyping and patient stratification. In the realm of biology and medicine, where large-scale multi-omics data, including genomics, transcriptomics, and epigenomics, is prevalent, deep learning models can enhance clustering algorithms. Their ability to reduce the dimensionality of complex data allows clustering algorithms to more effectively explore the heterogeneity between patients. Underscoring the utility of deep learning models, in particular autoencoders, in terms of data integration, dimensionality reduction, and handling a multitude of heterogeneous input data, Simidjievski et al. recently benchmarked various variational autoencoder models for multi-omics data [293].

Although patient stratification with deep learning methods are gaining traction in genomic data applications, they are often susceptible to external influences that are unrelated to the condition of interest. One severe limitation is the entanglement of biologically meaningful signals with variables unrelated to the inherent structure that one is interested in, i.e. technical artifacts, random noise from measurements, or other biological factors such as sex, ethnicity, and age (Figure 2.1a). These factors, referred to as confounders in the context of unsupervised learning, may cause clustering algorithms to form subgroups based on irrelevant signals, which may ultimately lead to spurious biological conclusions [294, 295].

Conventional strategies to account for and mitigate confounders involve training linear regression per feature against the confounder and take the residual part during pre-processing [296] or adjustments like pruning predictive dimensions after model training [297]. Conditional variational autoencoders (cVAE) have been used to create normative models considering confounding variables, such as age, for neurological disorders [298]. Dincer et al. proposed adversarial training to derive expression embeddings devoid of confounding effects [299], expanded upon by the single-cell Generative Adversarial Network (scGAN) for batch effect removal [300]. Liu et al. used a regularization term in the autoencoder's loss function to minimize correlation between latent embeddings and confounding bias [301]. Despite their methodological diversity, these methods have only been validated to work effectively on data from a single omics source and are not tailored towards disease subtyping and patient stratification.

To address this gap, we propose four novel VAE-based deconfounding frameworks for clustering of multi-omics data, utilising the i) removal of latent features correlated with confounders ii) a conditional variational autoencoder [298] iii) adversarial training [299,

300], and iv) adding a regularisation term to the loss function [301] as deconfounding strategies. To objectively assess whether our models can remove out-of-interest signals and find a patient clustering unbiased by confounding signals, we applied and evaluated our models on gene expression and DNA methylation pan-cancer data from The Cancer Genome Atlas (TCGA) program which we augmented with artificial confounding effects. In total, we simulated four different types of confounders, including linear, non-linear, categorical, and a mixture thereof, resembling realistic confounders such as age (linar, non-linear) [302–304] , BMI (non-linear) [305], or batch effects (categorical) [294, 299]. The contribution of our study is as follows:

- Four novel multi-omics clustering models based on VAE and different deconfounding strategies are presented.
- We highlight that various deconfounding techniques address confounded clustering in distinct ways, often overlooked within the algorithm's framework.
- Different confounding effects are simulated on the real-life TCGA dataset to demonstrate the influence of confounders on clustering and underscore the necessity for deconfounding models.
- Readers are provided with guidelines detailing strengths and limitations of each approach, along with suggestions on selecting an appropriate framework fitting their purposes.

5.2 Materials and methods

5.2.1 Data collection & preprocessing

This study utilized data collected within The Cancer Genome Atlas project (TCGA) [306], encompassing gene expressions (mRNA) of 4333 patients and DNA methylations (DNAm) of 2940 patients across six different cancer types, including breast invasive carcinoma (BRCA), thyroid carcinoma (THCA), bladder urothelial carcinoma (BLCA), lung squamous cell carcinoma (LUSC), head and neck squamous cell carcinoma (HNSC), and kidney renal clear cell carcinoma (KIRC). We prioritized these six types for their balanced sample sizes and excluded highly heterogeneous cancers to ensure robust clustering. Focusing solely on gene expression and DNA methylation omics, we aimed to optimize model performance and prevent overfitting. Additionally, selecting these omics introduced diversity in data

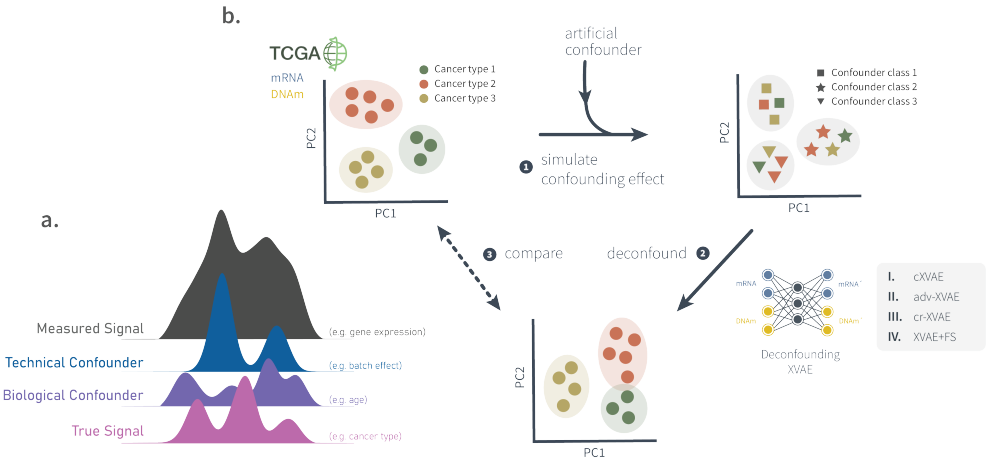


Figure 5.1: **a. A simplified graphical representation** of a measured signal (gray) which is a mix of independent sources such as the true signal (pink), a biological confounder (purple), and a technical confounder (blue). Note the difficulty of extracting the true signal from the measured additive signals. **b. Graphical summary of the work conducted in this study.** (1) Based on multi-omics pan-cancer TCGA data (section 5.2.1) different confounding effects were simulated (section 5.2.2). (2) Subsequently, four different deconfounding VAE frameworks (section 5.2.4) were trained on the the artificially confounded data. (3) The obtained deconfounded data was compared to the original, un-confounded input data in terms of clustering stability and deconfounding capabilities (section 5.2.6).

formats and distributions, with gene expression ranging continuously and DNA methylation exhibiting a largely bimodal beta distribution, enhancing the complexity and depth of our analysis.

Datasets were downloaded using the R package TCGAbiolinks [307]. The subsequent filtering step removed patients with (i) only a single data type available, (ii) missing clinical metadata, (iii) “american indian” or “alaska native” ancestry, and (iv) unknown tumor stage, resulting in a total of 2547 patients. The preprocessing of mRNA and DNAm data included the removal of probes (i) not shared across all cancer types, (ii) with missing values, and (iii) with 0 variance across all included patients, resulting in 58456 mRNA and 232088 DNAm features. To reduce the number of input features, we only considered the 2000 probes showing the largest variance across patients for each data type, resulting in a final data set of 2547 patients and 4000 features. This reduction strikes a balance between the number of features included and biological variability addressed and is in line with other clustering works on TCGA data [308]. TCGA-BRCA molecular subtype information for 724 (out of 731) patients was derived from the TCGA Pan-Cancer Atlas [309] *via* the cBioPortal [310].

5.2.2 Simulation of confounders

To imitate common confounding scenarios in real-life clustering applications we simulated linear, squared, categorical confounders, and a mixture thereof, resembling e.g. ageing [302–304], BMI [305], or batch effects [294, 299]. These confounders hinder the true or biologically meaningful clustering by intrinsically affecting the data structure in an unwanted way and possibly leading to a confounded clustering.
Here we denote the mRNA data as $X_1 \in \mathbb{R}^{n \times p}$ and DNAm data as $X_2 \in \mathbb{R}^{n \times q}$, where n, p, q are the number of patients, gene expressions, and DNA methylations, respectively. We first rescaled every mRNA and DNAm feature to the range $[0, 1]$ to avoid large ratios between the raw feature and the confounding effect. A visualisation of all confounding effects can be found in the Supplementary Methods.

Linear confounder

We uniformly generated a random numeric confounder $\mathbf{c} \in \mathbb{R}^n$ with discrete values $\{0, 1, 2, 3, 4, 5\}$, leading to a confounder clustering of six classes. Its linear effect on each individual is $\mathbf{c} + 5$ and a random weight for each feature was multiplied with it:

$$X_1' = X_1 + E_1^{\text{linear}} = X_1 + (\mathbf{c} + 5) \otimes \mathbf{w_1} \tag{5.1}$$

$$X_2' = X_2 + E_2^{\text{linear}} = X_2 + (\mathbf{c} + 5) \otimes \mathbf{w_2} \tag{5.2}$$

, where $\otimes$ denotes the outer product between two vectors, and $\mathbf{w_1} \in \mathbb{R}^p \sim U(0, 0.1)$, $\mathbf{w_2} \in \mathbb{R}^q \sim U(0, 0.2)$. We chose the uniform distribution of $\mathbf{w_1}$ to range from 0 to 0.1 so that the total linear effect would range from 0 to 1, having the same scale as X_1. We increased the upper bound of $\mathbf{w_2}$ to 0.2 due to our observation that X_2 is less sensitive to linear confounders.

Non-linear confounder

Non-linear effects were simulated in a similar way to linear effects. However, to mimic a non-linear confounder, as observed in, e.g. the significant quadratic association between body mass index and colon cancer risk [305], we considered adding an element-wise squared confounding effect $\mathbf{c}^2$ on the features:

$$X_1' = X_1 + E_1^{\text{square}} = X_1 + \mathbf{c}^2 \otimes \mathbf{w_1} \tag{5.3}$$

$$X_2' = X_2 + E_2^{\text{square}} = X_2 + \mathbf{c}^2 \otimes \mathbf{w_2} \tag{5.4}$$

, where $\mathbf{w_1} \sim U(0, 0.04)$, $\mathbf{w_2} \sim U(0, 0.04)$. The distribution of $\mathbf{w_1}$ and $\mathbf{w_2}$ was also determined based on the scale of X_1 and X_2.

Categorical confounder

The categorical confounding effect was achieved by shifting patients with the same confounder class to a distinctive direction in the feature space. More specifically, we first sampled six p-dimensional vectors for shifting mRNA data and six q-dimensional vectors for shifting DNAm data, both from $U(0, 1)$ and corresponding to six different confounder classes. The n patients were randomly assigned to each of the six categories. As a result, two matrices $C_1 \in \mathbb{R}^{n \times p}$ and $C_2 \in \mathbb{R}^{n \times q}$ denote the concatenation of shifting vectors of every patient for mRNA and DNAm, respectively. The categorical confounder is therefore the membership of all individuals in the six classes. A typical example of categorical confounders for clustering could be batch effects caused by collecting data from different centers [294, 299]. Then the confounded features were created via:

$$X_1' = X_1 + E_1^{\text{categ}} = X_1 + \text{diag}(\mathbf{w}) \cdot C_1 \tag{5.5}$$

$$X_2' = X_2 + E_2^{\text{categ}} = X_2 + \text{diag}(\mathbf{w}) \cdot C_2 \tag{5.6}$$

, where diag$(\cdot)$ converts a vector into its corresponding diagonal matrix. Different from the case of a numeric confounder, the weight vector $\mathbf{w} \in \mathbb{R}^n \sim U(0, 1)$ of the categorical confounder indicates to what extent every patient was shifted so that patients would have various strength of association with their confounder class.

Mixed confounder types

Real-life data analyses are likely affected by multiple confounders of different kinds, for instance, many cancer studies correct for age, age squared, education, etc. jointly in their models [303, 304]. Here we simulated a mixed confounding effect of linear, non-linear, and categorical confounders as described below:

$$X_1' = X_1 + E_1^{\text{linear}} + E_1^{\text{square}} + E_1^{\text{categ}} \tag{5.7}$$

$$X_2' = X_2 + E_2^{\text{linear}} + E_2^{\text{square}} + E_2^{\text{categ}} \tag{5.8}$$

, where $E_1^{\text{linear}}, E_2^{\text{linear}}, E_1^{\text{square}}, E_2^{\text{square}}, E_1^{\text{categ}}, E_2^{\text{categ}}$ represent the second term in Formula (1-6), respectively.

5.2.3 Variational autoencoder for data integration (XVAE)

A variety of different VAE architectures exist for the purpose of data integration, as extensively compared by Simidjievski et al. [293]. In this study, we utilize one architecture recommended by the respective authors, namely the X-shaped Variational Autoencoder (XVAE) (Figure 5.2). This architecture merges the heterogeneous input data sources into a combined latent representation z by learning to reconstruct each source individually from the common representation. Here we consider only two data types of a single datapoint x_1 and x_2, and the loss function of XVAE is as follow:

$$L_{\text{XVAE}}(\phi, \theta; x_1, x_2) = -E_{z \sim q_\phi(z|x_1,x_2)}[\log p_\theta(x_1, x_2|z)] + \beta * \text{MMD}(q_\phi(z|x_1, x_2)||p(z)) \quad (5.9)$$

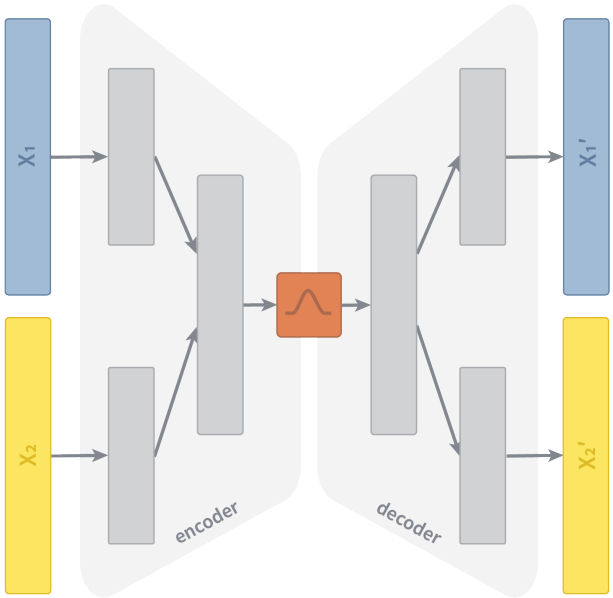


Figure 5.2: **Schematic representation of an X-shaped Variational Autoencoder (XVAE)**. The two input layers (X_1, X_2) denote the two omics dimension used in this study, namely gene expression and DNA methylation. The encoder consists of contiguous hidden layers, each with fewer nodes. We design the encoder of XVAE with a total of 2 layers prior to the latent embedding. In the first hidden layer, the dimension of each input entity is reduced individually. In the second hidden layer, input entities get fused into a combined layer. The latent embedding (red) represents the bottleneck of the XVAE with the minimum number of nodes. The decoder reversely mirrors the layer structure of the encoder, with the final layer featuring the same number of nodes as the input layer as it attempts to reconstruct (X_1', X_2') the original input from the latent embedding.

, where $q_\phi(z|x_1, x_2)$ encodes the latent space as a probability distribution over the input

variables (parameterised by ϕ) and $p_\theta(x_1, x_2|z)$ encodes the reconstruction of input variables as a probability distribution over the latent space (parameterised by θ). Following the originally proposed implementation, we use maximum mean discrepancy (MMD) as a regularization term to constrain the latent distribution q_ϕ to be a standard Gaussian distribution, balanced by the constant beta (β), which is set to 1 for all experiments. A more detailed description on autoencoders, as well as the XVAE architecture and training procedure can be found in the Supplementary Methods.

5.2.4 Multi-omics deconfounding models

Here, we will first describe in section 5.2.4 the use of linear regression for confounder correction and PCA for dimensionality reduction, which we deem the "baseline model" due to their wide popularity. Then, we outline in section 5.2.4 - 5.2.4 the four XVAE-based deconfounding models proposed in this study. Throughout this section we denote the confounder value of a single data point as c.

Baseline model: linear regression and PCA (LR+PCA)

Under the assumption that the effects of one or multiple confounders are linearly additive to the true signal of a feature, we build a linear regression (LR) model for the confounders against each mRNA or DNAm feature and then take their residuals as adjusted features. Subsequently, the adjusted features from the two data types are concatenated and their dimensionality is reduced via PCA (LR+PCA). We select the top 50 PCs explaining most of the variance of data to keep the embedding size identical to that of every XVAE-based model. The 50 PCs explaining the most variance of data are considered for the final clustering, for which KMeans with 10 random initialisations is applied.

Conditional X-shaped Variational Autoencoder (cXVAE)

Conditional variational autoencoder (cVAE) [311] is a semi-supervised variation of VAE, which originally aims to fit the distribution of the high-dimensional output as a generative model conditioned on auxiliary variables. Lawry et al. proposed to achieve deconfounding through a cVAE incorporating confounding variable information as auxiliary variables [298]. We extend this initial idea to be able to handle multi-omics data by replacing the originally proposed VAE with the XVAE model, resulting in a conditional X-shaped variational autoencoder (cXVAE) architecture (Figure 5.3A). We tested the integration of confounders at different levels of the cXAVE, including the input layer, the hidden layer

that fuses multiple inputs, and the embedding. More details on cXVAE implementations can be found in the Supplementary Methods.

X-shaped Variational Autoencoder with adversarial training (adv-XVAE)

The adversarial deconfounding autoencoder proposed by Dincer et al. [299] follows the idea of training two networks simultaneously - an autoencoder to generate a low dimensional embedding and an adversary multi-layer perceptron (MLP) trained to predict the confounder from said embedding (Figure 5.3B). By adversarially training the two networks, i.e. the autoencoder aims to generate an embedding which can not be used for confounder prediction by the MLP, it aims at generating embeddings that can encode biological information without encoding any confounding signal. As the original framework can only handle a single data type, we adapt it to work with multi-omics input by replacing its autoencoder with XVAE architecture. Details on architecture and training procedure of adv-XVAE can be found in Supplementary Methods.

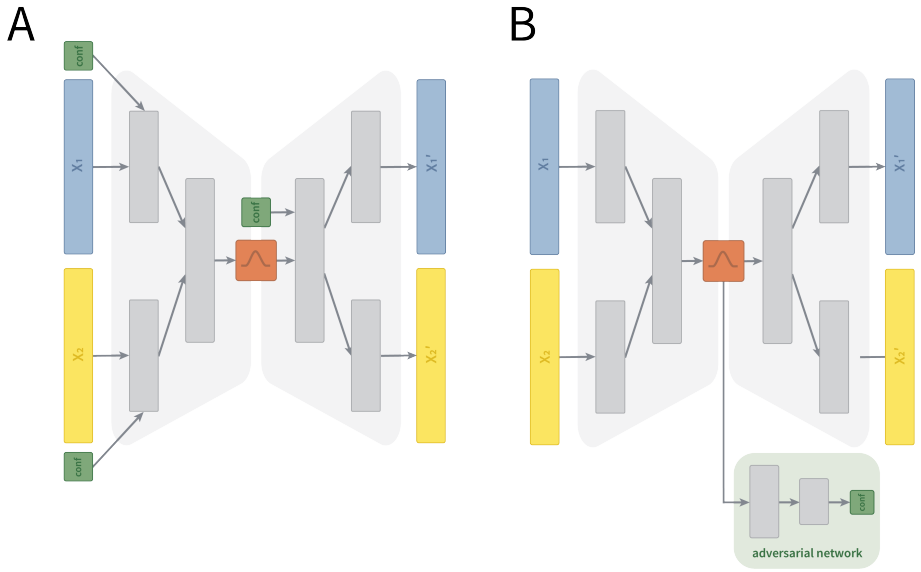


Figure 5.3: **Schematic representation of** (A) **conditional variational autoencoder (cVAE)** and (B) **adversarial deconfounding XVAE (adv-XVAE)**. (A) Depicts the cXVAE implementation termed *input + embed* due to the addition of confounders (green) in the first layer of the encoder and decoder. (B) Depicts the adv-XVAE implementation termed *multiclass* due to the usage of only a single supervised adverarial network (light green) trained to predict confounders (green) using a multiclass prediction loss. X_1 and X_2 are the two omics dimensions, namely gene expression and DNA methylation, while $X_1{}'$ and $X_2{}'$ denote their respective reconstruction. More details and visualisations of other implementation can be found in the Supplementary Material.

X-shaped Variational Autoencoder with deconfounding regularization (cr-XVAE)

Augmenting the loss function of deep learning models is an effective way to impose restrictions on the model or enforce learning of specific patterns. As an example, studies focused on disentangling the often highly correlated latent space of autoencoders impose constraints on the correlation between latent features by adding a penalty term to the loss function [301]. Inspired by this idea, we formulate a deconfounding regularization term aiming to reduce the degree of correlation between latent features and confounders. The regularized loss function becomes:

$$L_{\text{cr-XVAE}}(\phi, \theta; x_1, x_2, c) = L_{\text{XVAE}}(\phi, \theta; x_1, x_2) + f(z, c) \tag{5.10}$$

, where $f(z, c)$ denotes the joint association between latent features and confounders. More specifically, we choose two different association measurements, Pearson correlation and mutual information. Because Pearson correlation ranges from -1 (negatively correlated) to 1 (positively correlated) and both indicate strong relationship, we regularize only the magnitude of correlation by two methods, taking its absolute value or squared value. Because the confounder distribution needed for mutual information is usually unknown, we implement two methods to approximately compute mutual information as loss function, with differentiable histogram or kernel density estimate.

Feature selection by removing correlated latent features (XVAE+FS)

The removal of latent features correlated with confounders comes from the idea of *post hoc* interpretation of latent features [312]. To identify confounded latent features, we calculate the Pearson correlation between each latent variable and the confounder. For determining the threshold indicating which latent features are being removed from further analyses, we test two different approaches:

1. *p-value cutoff* - the p-value of the Pearson correlation indicates the probability that the computed correlation is smaller than a random correlation between uncorrelated datasets. Latent features with a p-value < 0.05 are excluded from analyses.

2. *absolute correlation coefficient* - Pearson correlation measures the linear relationship between two variables. Latent features exhibiting an absolute Pearson correlation of more than 0.3 (weak correlation) or 0.5 (strong correlation) are excluded.

5.2.5 Consensus clustering

Different from the baseline linear regression model which adopts KMeans on the deconfounded features for clustering, we apply consensus clustering on the latent features of each VAE-based deconfounding model. Here, consensus clustering takes the advantage of random sampling in a VAE and it aggregates the individual clustering of each embedding sampled from the latent distributions [313]. We first generate 50 embedding matrices for all the n samples, on each of which a KMeans clustering is performed. Subsequently, a consensus matrix $\bar{A} \in \mathbb{R}^{n \times n}$ is constructed from all the 50 clusterings:

$$\bar{A} = \frac{1}{50} \sum_{i=1}^{50} A_i \tag{5.11}$$

, where $A_i \in \mathbb{R}^{n \times n}$ is the binary matrix of each KMeans clustering indicating if two data points are assigned to the same cluster or not. Values of $\bar{A}$ are in the range [0,1], where 0 means the two corresponding samples are never clustered together in the 50 clusterings while 1 means they are always in the same cluster. Finally, a spectral clustering is performed on the consensus matrix $\bar{A}$ to derive a stable clustering of the patients. To experiment on the potential impact of the number of embedding matrices, we rerun the model with various numbers (10, 50, 100, 150, 200) and compare the model performance (Supplementary Table 4).

5.2.6 Evaluation metrics

We apply each of the aforementioned models to the artifically confounded multi-omics dataset described in section 5.2.1 and 5.2.2. Every model is evaluated in terms of their XVAE reconstruction accuracy, measured as the relative reconstruction error of inputs, their clustering stability, evaluated by the dispersion score of consensus clustering (CC), and deconfounding capabilities for clustering, estimated by calculating the Adjusted Rand index (ARI) for true (cancer types) and confounder labels.

XVAE reconstruction accuracy

Model training is monitored through inspection of the validation loss. To evaluate reconstruction quality of the trained XVAE model, we compute the L2 relative error (RE) between the original input (x) and reconstructed data ($x^{'}$) for (i) each data type individually:

$$\mathrm{RE} = \frac{\sqrt{\sum_{i=1}^{n} \|x_i - x_i^{'}\|^2}}{\sqrt{\sum_{i=1}^{n} \|x_i\|^2}} \tag{5.12}$$

, as well as (ii) for the combined data types:

$$\mathrm{RE} = \frac{\sum_{m=1}^{2} \sqrt{\sum_{i=1}^{n} \|x_{mi} - x'_{mi}\|^2}}{\sum_{m=1}^{2} \sqrt{\sum_{i=1}^{n} \|x_{mi}\|^2}} \tag{5.13}$$

, where $m = 1, 2$ indicates the two data types.

Clustering stability

Before assessing how well each model can derive a meaningful clustering, we want to first check if a model can stably cluster the samples. To achieve this goal, we employ the dispersion score to measure the internal stability of consensus clustering based on its consensus matrix $\bar{A}$:

$$\mathrm{Dispersion} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (\bar{A}_{ij} - 0.5)^2 * 4}{n^2} \tag{5.14}$$

The dispersion score ranges from 0 to 1, where 1 shows a perfect stability that every value in $\bar{A}$ is either 0 or 1, i.e. no confusion among the clusterings, and the lower the less consensus among the clusterings.

Deconfounding capabilities

We compare our clustering with two different labels, the true one, namely cancer types, and the confounder. An ideal model should deconfound the features sufficiently while keeping the meaningful information for obtaining the true clustering. In other words, we expect a model with high ARI with the true label and low ARI with the confounder label. The association between true patient label and clusters obtained when modelling the original (unconfounded) data represents the best achievable clustering, with ARI value converging towards 1.

Similar to ARI, we compute another external clustering metric, the normalized mutual information (NMI), which measures the dependence between two clusterings. As it only shows complementary information to ARI, we record the NMI of every clustering model in Supplementary Table 1.

5.2.7 Implementation

For better stability and generalization, we train each model 50 times using i) randomly sampled training and validation sets with a ratio of 80:20 and ii) different seed of randomization.

5.2.8 Software

All of the models described in this study are built in Pytorch Lightning [314] and trained using the GPU units RTX 2080 Ti 11GB.

5.3 Results

5.3.1 cXVAE outperforms other considered deconfounding strategies in the presence of a single confounder

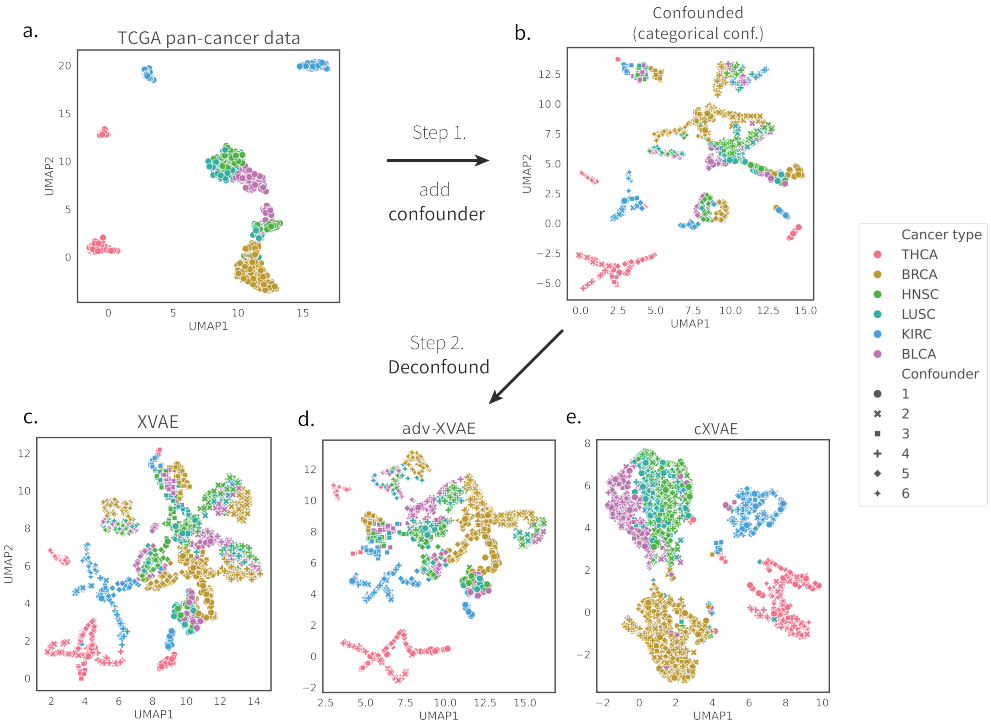


Figure 5.4: **Deconfounding behaviour of several developed models**. Dimensionality reduction (UMAP) plot of the (a.) unconfounded TCGA pan-cancer data, (b.) categorically confounded data, as well as deconfounding using the (c.) vanilla XVAE (d.) adv-XVAE, and (e.) cXVAE. Marker colors indicate the true label labels (i.e. TCGA cancer types), while marker shapes indicate the six classes (1-6) of the confounder (see section 5.2.2). The steps indicated describe the experimental order (see Figure 2.1).

We simulated different types of confounding effects - linear, non-linear (squared), and categorical - on the multi-omics TCGA pan-cancer dataset to benchmark a total of four deconfounding frameworks, namely XVAE with Pearson correlation feature selection (XVAE+FS),

conditional XVAE (cXVAE), adversarial training with XVAE (adv-XVAE), and confounder-regularised XVAE (cr-XAVE) (see Methods for more details). We additionally included two baseline models to compare with: 1) confounder correction with linear regression (LR+PCA) and 2) vanilla XVAE without any deconfounding (XVAE). To estimate the robustness of each method, each model was trained on 50 iterations of randomly sampled training and validation data (80:20 split) and random seed initialization.

All proposed deconfounding approaches were able to correct for a linear confounder, as denoted by the high ARI for true clustering and low ARI for confounder clustering (Table 5.1). Performances started to decline for non-linear confounding problems, with cXVAE clearly outperforming other strategies. For non-linear confounders we noted large ARI for confounder clustering across all strategies and simulation setups. This illustrates that, while good clustering performance for true labels were achieved, the full removal of unwanted signal was not easily achievable for all the models. Categorical confounding was perceived to be the most difficult, with all models except cXVAE exhibiting a high decrease in true clustering performance. Notably, cr-XVAE and XVAE+FS were able to remove artificial confounders completely, however at the cost of simultaneously removing true clustering signal. adv-XVAE, which in theory should be a strategy well suited to deal with categorical problems, fails to consistently remove the categorical confounding effect. In general we noted a decline of reconstruction accuracy of models with increasing complexity of the confounder simulations.

To illustrate the deconfounding capabilities of several of the developed models, we examined their clustering results obtained on the TCGA dataset involving categorical confounders (Figure 5.4). The UMAP plot of the original, unconfounded data (Figure 5.4, a.) shows the distinct clustering of THCA (thyroid carcinoma), KIRC (kidney renal clear cell carcinoma), and BRCA (breast invasive carcinoma), while the clustering of BLCA (bladder urothelial carcinoma), LUSC (lung squamous cell carcinoma), and HNSC (head and neck squamous cell carcinoma) appears to be more entangled. The observed clustering was significantly obscured by the addition of an artificial categorical confounder, visible as the clustering being dictated by the confounder class rather than cancer type (Figure 5.4, b.). An attempted deconfounding using a vanilla XVAE (Figure 5.4, c.) or adv-XVAE (Figure 5.4, d.) model displayed little improvement over the confounded clustering, demonstrating the models' inability to remove the artificially introduced signal. The cXVAE model, however, proved to be able to effectively mitigate the confounding effect, resulting in a clustering similar to the original, unconfounded data (Figure 5.4, e.). In an attempt to investigate whether the deconfounding using cXVAE not only restores pan-cancer type but can also recover cancer subtypes, we examined the clustering of several TCGA-BRCA

molecular subtypes, including Her2, LumA, Basal, LumB, and normal, before and after deconfounding (Supplementary Figure 5). This revealed that while molecular subtypes were completely masked by the simulated categorical confounders (Supplementary Figure 5, b), deconfounding with cXVAE could retrieve their original clustering (Supplementary Figure 5, c).

In summary, across all confounder simulations, cXVAE clearly outperformed other deconfounding strategies in terms of clustering accuracy, deconfounding power, and model robustness. The ARI on true clustering obtained by cXVAE in all three scenarios reached around 0.7, which is very close to the performance of the vanilla XVAE on unconfounded data (0.731, see details in Supplementary Table 2).

A more detailed summary of the performances of each model can be found in Supplementary Table 1.

The dimensionality reduction plots for models not displayed in Figure 5.4, namely LR+PCA, XVAE+FS, and cr-XVAE can be found in Supplementary Figure 2. It illustrates that while XVAE+FS and cr-XVAE yield performances similar to XVAE and adv-XVAE, LR+PCA seems unable to distinguish the true signal from the confounder signal. Furthermore, the deconfounded clustering derived by cXVAE in the presence of multiple confounder is shown in Supplementary Figure 4, which indicates the capabilities of cXVAE to deconfound even in this complex scenario. While Table 5.1 depicts the best performing implementation of each deconfounding model, we tested a number of possible implementations (see Methods), which we observed to have a notable impact on model performance (Supplementary Table 2). Therefore, we provide design recommendations for each deconfounding strategy in the Supplementary Results.

5.3.2 cXVAE is easily extendable to handle multiple confounders of mixed types

In a realistic setting datasets can be confounded by multiple confounders with different biasing effects. In an attempt to investigate how well deconfounding strategies can handle more than one confounder, we simulated the parallel presence of three confounders of different effect, namely linear, non-linear, and categorical (Table 5.2). In line with our observations with the single confounder simulations, cXVAE outperformed other models in terms of true clustering accuracy and deconfounding efficiency. While also other strategies like XVAE+FS, cr-XVAE, or LR+PCA were able to successfully remove all three simulated effects, they achieved this at the cost of true signal. adv-XVAE failed to fully remove confounders, while also showing very low true clustering accuracy and can therefore be considered unsuitable for the task. We also noted that the decline in reconstruction

accuracy with increasingly complex confounding situations is even more pronounced in multiple confounder settings.

5.3.3 cXVAE is able to retrieve biology-driven clustering from confounded data

To illustrate the deconfounding capabilities of cXVAE, the model that outperformed others across all four evaluation metrics in various confounding scenarios, we examined the clustering results obtained on the TCGA dataset involving categorical confounders (Figure 5.4). The UMAP plot of latent features clearly showed that BRCA (breast invasive carcinoma), THCA (thyroid carcinoma), and KIRC (kidney renal clear cell carcinoma) were well clustered by cXVAE, while BLCA (bladder urothelial carcinoma), LUSC (lung squamous cell carcinoma), and HNSC (head and neck squamous cell carcinoma) were still entangled. In summary, we found the deconfounding behaviour of cXVAE to not only yield clusterings resembling those of the unconfounded data, but also be in line with the pathological and physiological differences between the pan-cancer types. BLCA arises from urothelial cells in the transitional epithelium, which can change from cuboidal to squamous form when stretched. Furthermore, squamous differentiation is by far the most common histological variant of urothelial carcinoma [315], indicating a close relationship between urothelial carcinoma and squamous cell carcinoma. Apart from BLCA, the overlap in clustering of LUSC and HNSC can be directly explained by their common origin of squamous cells, while BRCA, THCA, and KIRC are all carcinoma related to glandular cells [316]. Supporting the validity of our obtained cXVAE clustering, other multi-omics pan-cancer studies utilising stacked variational autoencoders [317], penalized matrix factorization [318], or supervised VAE [319] have retrieved similar cancer type clustering.

5.4 Discussion

In this study, we addressed the possible harm of ignoring or inadequately handling confounders to clustering samples with (multi-)omics measurements. In epidemiology, a confounder is a variable that can affect the result of a study because it is related to both the exposure and the outcome being studied. Here, we extended the definition to unsupervised models for disease subtyping to indicate variables that can distort the relationship between inferred or predicted cluster membership and disease.
Extensive simulation revealed that cXVAE stands out as a versatile and accurate deconfounding approach. The applicability of conditional autoencoder to biological data to e.g.

correct for batch effects [298] or disentangle confounders in fMRI [320] or microRNA data [321] has been shown before. However, by merging the principles of a conditional autoencoder with the framework of an autoencoder specifically tailored for the integration of multi-omics data, our research charts new frontiers in the domain of deconfounded patient stratification.

While adversarial training may offer an alternative flexible deconfounding approach, we confirm that optimization of model hyper-parameters is challenging [300]. Instability may become more pronounced in the presence of multiple confounders. This can be explained by the fact that adversarial networks were trained separately for each confounder, sequentially adding extra terms to its objective function (see Supplementary Table 2).

In the literature, a statistical correlation loss has been proposed to replace the adversarial prediction loss in a adversarial training model [322], resembling our cr-XVAE model. The difference is that cr-XVAE directly computes the correlation between the VAE embedding and the confounder without an additional adversarial network. We implemented Pearson correlation and mutual information as the regularization term of cr-XVAE but other association measures could also be adopted, e.g. Spearman correlation and cosine similarity. In the case of multiple confounders, it is also possible to weigh their associations differently in the loss function to balance deconfounding strength.

The identification of disease subtypes requires performing a clustering algorithm at some point. Even though iterative training of the clustering in a joint autoencoder loss function can overcome inconsistencies between training and downstream clustering performance [323–325], we chose for a decoupled strategy. This was to 1) avoid having too many terms in loss function to confuse training, and 2) reduce computation time and initialization settings with iteratively training clustering in a joint loss function. Consensus clustering furthermore has several advantages in data science including robustness, stability, interpretability and flexibility, as it can be applied to various types of data and clustering algorithms. Our consensus clustering scheme adopts spectral clustering as its final step because the consensus matrix can be naturally viewed as a graph of all patients and the superior performance of spectral clustering has been shown on graphs [326]. We chose to sample and cluster the embedding for 50 iterations to construct the consensus matrix. Having more iterations can improve the robustness of consensus clustering but at the cost of computation time. We observed that the 50 individual clusterings are very consistent and increasing the number of iterations won't necessarily improve the final performance (Supplementary Table 4).

It remains a daunting task to generate data that adequately reflect the complexity of real-life cases. Therefore, one needs to be aware that simulations of confounders always

represent simplifications of real observable effects. Note that the range of weight vectors $\mathbf{w_1}$ and $\mathbf{w_2}$ may have an important influence on how the data is confounded. A large weight will cause a stronger confounding effect and potentially increase difficulty in finding the true clustering. Currently, we set their values based on the scale of the original features to balance between the true signal and confounder signal. While this study is limited to the use of two data types, in principle the XVAE design utilised allows the integration of heterogeneous data from many more sources simultaneously. Additionally, since all evaluated deconfounding strategies share the same XVAE design as a foundation, we anticipate consistent training time and performance across models when scaling up dimensions.

5.5 Conclusion

In this study, we presented four VAE-based multi-omics clustering models and their variations, following different deconfounding strategies. Their clustering and deconfounding performance was evaluated and compared with baseline models on the multi-omics pan-cancer dataset from TCGA with artificailly generated confounding effects. The results showed both the necessity to adjust for confounders and that our novel models, cXVAE in particular, can effectively deal with the confounding effects and obtain the biologically meaningful clustering. We demonstrate that our multi-omics deconfounding VAE clustering models have big potential in delivering accurate patient subgrouping or disesase subtyping, ultimately enabling better personalised healthcare.

linear

	Reconstruction error	CC dispersion	Clustering performance (ARI)	
			True label (cancer type)	Confounder
LR+PCA	-	-	0.692	0.001
XVAE	0.246 (± 0.004)	0.844 (± 0.045)	0.506 (± 0.116)	0.151 (± 0.057)
XVAE+FS	0.245 (± 0.004)	0.860 (± 0.028)	0.571 (± 0.092)	0.008 (± 0.007)
cXVAE	**0.234 (± 0.003)**	**0.935 (± 0.023)**	**0.712 (±0.055)**	**0.001 (± 0.001)**
adv-XVAE	0.245 (± 0.004)	0.901 (± 0.032)	0.568 (± 0.070)	0.093 (± 0.051)
cr-XVAE	0.244 (± 0.003)	0.873 (± 0.028)	0.598 (± 0.074)	0.004 (± 0.004)

non-linear

	Reconstruction error	CC dispersion	Clustering performance (ARI)	
			True label (cancer type)	Confounder
LR+PCA	-	-	0.391	0.215
XVAE	0.236 (± 0.003)	0.826 (± 0.042)	0.307 (± 0.141)	0.297 (± 0.071)
XVAE+FS	0.236 (± 0.003)	0.805 (± 0.040)	0.411 (± 0.142)	0.138 (±0.078)
cXVAE	**0.227 (± 0.002)**	**0.908 (± 0.031)**	**0.646 (± 0.079)**	**0.076 (± 0.074)**
adv-XVAE	0.238 (± 0.004)	0.942 (± 0.025)	0.568 (± 0.049)	0.194 (± 0.006)
cr-XVAE	0.235 (± 0.002)	0.852 (± 0.043)	0.478 (± 0.129)	0.154 (± 0.042)

categorical

	Reconstruction error	CC dispersion	Clustering performance (ARI)	
			True label (cancer type)	Confounder
LR+PCA	-	-	0.150	0.071
XVAE	0.216 (± 0.003)	0.762 (± 0.055)	0.330 (± 0.125)	0.048 (± 0.088)
XVAE+FS	0.216 (± 0.003)	0.787 (± 0.040)	0.361 (± 0.100)	0.010 (± 0.023)
cXVAE	**0.210 (± 0.002)**	**0.911 (± 0.033)**	**0.664 (± 0.070)**	**0.001 (± 0.001)**
adv-XVAE	0.217 (± 0.002)	0.764 (± 0.058)	0.240 (± 0.188)	0.156 (± 0.084)
cr-XVAE	0.216 (± 0.003)	0.813 (± 0.034)	0.368 (± 0.101)	0.001 (± 0.001)

Table 5.1: **Overview performances of deconfounding strategy for single confounder simulations**. Values are displayed as mean ± standard deviation of 50 runs with different parameter initialisation and randomly sampled training and validation data. Models on the first column indicate the following deconfounding strategies and implementations thereof: linear regression followed by principal component analysis and KMeans clustering (LR+PCA), vanilla XVAE without any deconfounding (XVAE), XVAE with feature selection in the form of removing correlated latent features (XVAE+FS, correlation cutoff = 0.5), conditional XVAE (cXVAE, input + embedding), adversarial training with XVAE (adv-XVAE, multiclass MLP), confounder-regularised XVAE (cr-XAVE, squared correlation regularisation). Reconstruction error: relative error in the reconstruction of X1 and X2 weighted eqally; CC dispersion: consensus clustering agreement over 50 iterations; True clustering: adjusted rand index (ARI) of consensus clustering derived clusters with True label labels; Confounder clustering: ARI of consensus clustering derived clusters with simulated confounder labels.

	multiple confounders					
			Clustering performance (ARI)			
	Reconstruction error	CC dispersion	True label (cancer type)	Linear confounder	Squared confounder	Categorical confounder
LR+PCA	-	-	0.215	0.001	0.014	0.001
XVAE	0.161 (± 0.003)	0.725 (± 0.043)	0.216 (± 0.089)	0.015 (± 0.019)	0.140 (± 0.043)	0.067 (± 0.048)
XVAE+FS	0.161 (± 0.003)	0.731 (± 0.037)	0.265 (± 0.085)	0.007 (± 0.009)	0.019 (± 0.030)	0.109 (± 0.057)
cXVAE	**0.146 (± 0.002)**	**0.905 (± 0.022)**	**0.634 (± 0.042)**	**0.001 (± 0.001)**	**0.001 (± 0.001)**	**0.001 (± 0.001)**
adv-XVAE	0.158 (± 0.004)	0.753 (± 0.066)	0.225 (± 0.120)	0.016 (± 0.023)	0.107 (± 0.052)	0.106 (± 0.051)
cr-XVAE	0.161 (± 0.003)	0.764 (± 0.031)	0.369 (± 0.064)	0.003 (± 0.002)	0.007 (± 0.010)	0.001 (± 0.001)

Table 5.2: **Overview performances of deconfounding strategy in the presence of multiple confounders**. Values are displayed as mean ± standard deviation of 50 runs with different parameter initialisation and randomly sampled training and validation data. For a detailed description of columns and models, please refer to Table 5.1.

5.6 Supplementary Notes

All Supplementary Material, including figures and tables, are available under: https://doi.org/10.1101/2024.02.05.578873

5.6.1 Data and code availability

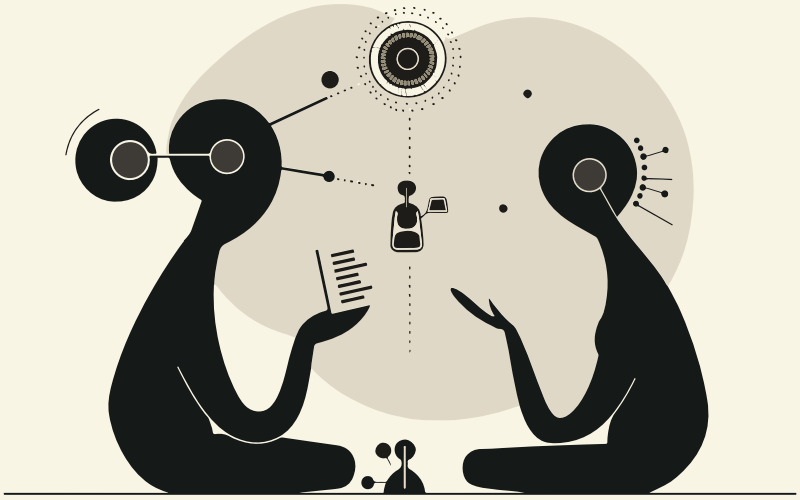
Our code is available at https://github.com/ZuqiLi/Multi-view-Deconfounding-VAE. The simulated data generated in the course of this study are available at Zenodo (https://doi.org/10.5281/zenodo.10458941).

5.6.2 Acknowledgements

The authors thank the supporters of this study, namely the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860895 TranSYS. Furthermore we thank the members of the Computational Population Biology group at Erasmus Medical Center for their critical and creative input to this work. We also would like to extend our gratitude to the PhD candidates enrolled in the "*Frontières de l'Innovation en Recherche et Éducation*" (FIRE) doctoral school and thank them for their critical reviewing and feedback on the pre-print of this study.

5.6.3 Funding

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement [860895 to Z.L., S.K. and K.V.S.]; E. S. acknowledges the funding received 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 personalised medicine-smart combination of pre-clinical and clinical research with data and ICT solutions).



Part 3

Model Explainability

Chapter 6

Designing Interpretable Deep Learning Applications for Functional Genomics: a Quantitative Analysis

Arno van Hilten[†], **Sonja Katz**[†], Edoardo Saccenti, Wiro J. Niessen, and Gennady V. Roshchupkin

[†]authors contributed equally

Manuscript submitted

Abstract

Deep learning applications have had a profound impact in many scientific fields, including functional genomics. Deep learning models can learn complex interactions between and within omics data, however interpreting and explaining these models can be challenging. Interpretability is not only essential to help progress our understanding of the biological mechanisms underlying traits and diseases but also for establishing trust in these models' efficacy for healthcare applications. Recognizing this importance, the recent years has seen the development of numerous diverse interpretability strategies, making it increasingly difficult to navigate the field. In this review, we present a quantitative analysis of the challenges arising when designing interpretable deep learning solutions in functional genomics. We explore design choices related to the characteristics of genomics data, the neural network architectures applied, and strategies for interpretation. By quantifying the current state of the field with a predefined set of criteria, we find the most frequent solutions, highlight exceptional examples, and identify unexplored opportunities for developing interpretable deep learning models in genomics.

6.1 Introduction

The overarching goal of functional genomics research is to understand and intervene in the underlying biological processes between genotype, environment and phenotype [327]. To probe these underlying biological processes, a wide variety of omics are gathered. Genetic data (genotype arrays, DNA sequences) serves as the stable foundation from which many biological processes start. Transciptomics data, provides the expression of genes, dynamically transcribed from the DNA it can be regulated by various epigenetic mechanism. These epigenetic mechanisms include, DNA methylation, RNA methylation, regulatory non-coding RNA, histone modifications and chromatin accessibility. All together, these omic types provide insight in the biological processes that link heritable and environmental factors to observable characteristic in a cell or individual [327]. These underlying biological processes can be very complex [328] and massive amounts of data are acquired to study these processes. The massive amount of data and the complexity of the biological processes, allow and justify the use of more complex models, such as deep learning models to help unravel the complexities between genotype, environment and phenotype.

Deep learning, a subset of machine learning, consists of a group of methods that can capture complex interactions and non-linear relationships given a sufficiently large number of examples. Deep learning solutions have had successes in a wide variety of applications, a few examples include: biomedical image segmentation [329] natural language modelling [330] and protein modelling [331]. Despite the widespread popularity and convincing performances of deep learning models there are drawbacks in using neural networks; deep learning methods offer no explanation for their decision-making processes.

Explaining the decision making process of AI-driven technologies is essential given that these technologies impact various facets of our daily lives, encompassing critical domains such as healthcare, governance, and legal systems [89]. Recent European data and privacy legislation, particularly the General Data Protection Regulation (GDPR), has put “explainability” as a top priority in machine learning research [109]. In the special case of decisions reached using automated processing, the right of data subjects (e.g. patients) were phrased as to "*obtain an explanation of the decision reached*" [332], encoding the right to explanations for data subjects within European law. Interpretability is central for inspiring trust in a neural network. Understanding *why* and *how* a model make decisions, as opposed to blindly trusting in that they are correct, contributes to a model’s trustworthiness [89]. The ability to understand the reasoning behind a model’s prediction (the *why*) is often termed "explainability", whereas the ability to fully understand the inner workings of the model (the *how*) is refereed to as its "interpretability". According to this

definition, all interpretable AI is inherently explainable but not all explainable AI systems are interpretable. Generally, rule-based models, and simpler machine learning models such as linear and logistic regression, decision trees are intrinsically interpretable. More complex models, like large random forests and dense neural networks, predictions can often only be approximated through explainability methods [90].

Designing an interpretable deep learning application is an inherently creative process without a set path to a successful application. Researchers require not only an in-depth knowledge of the data they are working with, but also need to be aware about the different types of model architectures that can be applied, how these architectures can be made interpretable, and finally combine it in a way that yields insight into the biological question of interest. Fortunately, there are numerous successful applications of interpretable deep learning in genomics that can guide development and inspire novel approaches. In this study, we provide an in-depth analysis of the current state of the interpretable genomics field by discussing the most prevalent solutions, visualize common combinations of solutions, and identify interesting opportunities for designing interpretable deep learning applications in genomics. While there exist numerous comprehensive reviews outlining different interpretability approaches [333–337] we provide a more practical guide for researchers that want to bring interpretability into their models. We dissect the challenges associated with the three key considerations of every interpretable deep learning application in genomics: (i) the characteristics of the utilised data, (ii) the model architecture, and (iii) the interpretation strategy selected. We assessed each surveyed paper according to a predefined set of criteria relating to these key considerations (Section 6.2), resulting in an overview table that provides a quantitative overview of all approaches and applications of interpretable deep learning in genomics (Section 6.3). From this comprehensive overview table, we extract general statistics that provide insight into the prevalent research questions addressed and the popularity of both models and interpretability methods. Furthermore, we provide supporting graphics to outline (dis-)advantages for combinations of data types, models, and interpretation strategies, with the goal to identify inspirational examples, challenges, and unexplored opportunities in developing an interpretable deep learning model.

In the first section **Considerations for Designing an Interpretable Model** we discuss the major decisions that need to be taken during model development and we provide a short intuition on the most commonly used data types, models, and interpretability strategies. After establishing the fundamentals, we analyze and visualize trends and statistics of current solutions in section **The current state of the field: a quantification**. Finally, in **Opportunities and perspectives**, we highlight unexplored opportunities, good practices

and considerations for designing interpretable deep learning applications in genomics.

6.2 Considerations for Designing an Interpretable Model

We focus on three key aspects that mark major decisions during developing an interpretable deep learning model for omics data: (i) the type and characteristics of data used, (ii) the model architecture of choice, and (iii) the interpretability strategy deployed. As all these three aspects are intertwined and can not be discussed separately, we aim on clarifying dependencies by gradually stacking information. First we expand and motivate the criteria for the quantitative analysis, before moving on to the quantitative analysis itself.
Criteria and considerations used for evaluation:

1. **Characteristic of the input data**
   (a) *Sequencing type*: was the data generated using single cell or bulk sequencing?
   (b) *Omic type*: which omics (e.g. SNPs, CNVs, mRNA, CpGs) were used in the study? Single omic or multi-omic?
   (c) *Data dimensions*: what is the number examples (e.g number of patients, cells) in the dataset?
2. **Choosing a model architecture**:
   (a) *Neural network type*: what kind of neural network architecture was used (CNN, VNN, transformer etc.)
   (b) *Input dimensions*: what was the dimension of the input for each example to the network? (e.g between 50 and 100, between 1000 and 2000, more than 1 million)
   (c) *Computational resources*: which computational resources (CPU/GPU memory) was available for model construction and interpretation?
3. **Navigating interpretation strategies**
   (a) *Biological level of interpretation*: on which (biological) level was interpretability applied (gene-level, pathway etc.)?
   (b) *Interpretation taxonomy*: how was model interpretation facilitated (global, local, attribution methods, hidden semantics etc.)?
   (c) *Interpretation strategy*: what are the defining characteristics of the interpretation method (use prior knowledge, visualisation, backpropagation method etc.)?

(d) *Prior knowledge*: If prior knowledge was used, which database was used? (KEGG, Reactome, Gene-Onthology etc.)

6.2.1 Characteristics of the input data

The data is the basis on which the neural networks are built and it is characteristics largely influence the choice of model architecture and the interpretation method. We included studies with at least one of the following types of data: genetic (SNPs, CNVs), transcriptomics (mRNA), epigenetic data such as chromatin accessibility (ATAC-seq, DNase), non-coding RNAs (ncRNAs) and DNA methylation (CpGs). We make a distinction between single-cell and bulk, as the challenges and characteristics between these sequencing types can be quite different and we categorized the number of examples in the dataset to provide an impression of the volume of the data needed to perform the study.
Table 6.1 highlights some of the characteristics of the omic types and challenges associated with designing a neural network for the included omic types. These characteristics and challenges, in combination with the research question mainly shape the realm of possible options for neural network architectures. For instance, due to the expansive dimensions of genetic data, utilizing sparse models becomes necessary, since fully connected layers could exceed GPU memory capacities. For sequence data, neural networks that "scan" the sequences for patterns, such as convolutional neural networks are typically utilized. However, the atypical out-of-the box applications, are interesting to highlight. These demonstrate that with conventional and unconventional transformations of the data, one can open up new possibilities. For example, the use of k-mers to decrease the input size of genetic data, and to increase the depth of the data [338]. Transforming the gene expression data to images in order to apply convolutional neural networks and image-based interpretation strategies [339–341]. ChromBPnet [342] avoids peak-calling by using the raw count an additional network to correct for a bias in the measurements of ATAC-seq (TN5 bias). This latter is also a great example for demonstrating that a thorough understanding of the data and prepossessing steps, is crucial in identifying steps that can be replaced or benefit from deep learning.

6.2.2 Choosing a model architecture

The choice of model architectures is mainly driven by data, technological innovations, and trends. Most deep learning architectures are variations of neural networks, we categorize each network as one of the main neural network types: multi-layer perceptron (MLP), convolutional neural networks (CNN), graph neural networks (GNN), autoencoders (AE),

	Genetic	Transcriptomics		Epigenetic	
Category	SNPs, sequences	Bulk gene expression	Single-cell gene expression	Chromatin accessibility	Methylation
Methods	Genotype arrays, whole genome sequencing	RNA-seq, microarrays	scRNA-seq	ATAC-seq, DNase-seq	BeadChip
Number of measurements	>88 million variants	~24,000 genes	>24,000 genes	Millions of reads	~450,000 CpGs
Data type	Categorical (nucleotide, dosage)	Positive continuous (gene expression level)	Positive continuous (gene expression level)	Positive continuous (read counts per region)	Fraction (CpG methylation level)
Challenges for ML	Large input size, small effect sizes, non-coding regions	Input order, mixture of cell signals	Identifying cell-types and cell-states, data sparsity	TN5 bias, peak-calling	High dimensionality, cell-type heterogeneity

Table 6.1: **Overview of the characteristics and challenges of the main omics types encountered in the surveyed papers.**

visible neural networks (VNN) and the recently introduced generative pretrained transformers (GPT). Historically, convolutional neural networks were designed for image data, graph neural networks for data that can be represented as graphs (e.g. social networks, molecules and proteins), and transformers had their first successes in text-based natural language tasks. However, neural networks can consist of a mix of multiple types of layers, providing endless opportunities to tailor the neural network to specific problems and data types, such as the various types of omics data. Genomics data generally has a large input dimension and one can choose to modify the network to take all the inputs or choose a subset of the data to feed into the network. This is related to the last criterion, the computational resources, larger networks that take all the data need more memory train longer.

Multi-layer perceptron (MLP) are the most traditional neural networks [344]. Each layer consist of a set of neurons that is fully connected to neurons in previous and subsequent layers (see Figure 6.1a). These feed-forward neural networks do not need to be deep with multiple subsequent layers to model complex functions. It has been shown that a shallow networks with a single layer, with infinite width, can accurately approximate any function [345, 346].

Convolutional neural networks (CNN) have a rich history of successful applications, particularly in imaging, where they excel in extracting useful patterns from local correlation structures, such as edges in images [344]. CNN are not fully connected, instead, in each convolutional layer it optimises a predefined number of filters that slide across the input features (as shown in Figure 6.1b). This sliding operation over the inputs makes the network invariant to where the pattern is located, in other words, the network is translation-invariant. Stacking multiple layers results in a fully convolutional neural network. In such a network, each subsequent layer can capture more abstract patterns. Stacking convolutional layers also contributes to increasing the receptive field; the region used by the network to create a particular feature. The receptive field defines thus the largest distance for which interactions can be learned by the network. In the context of genomics, this can

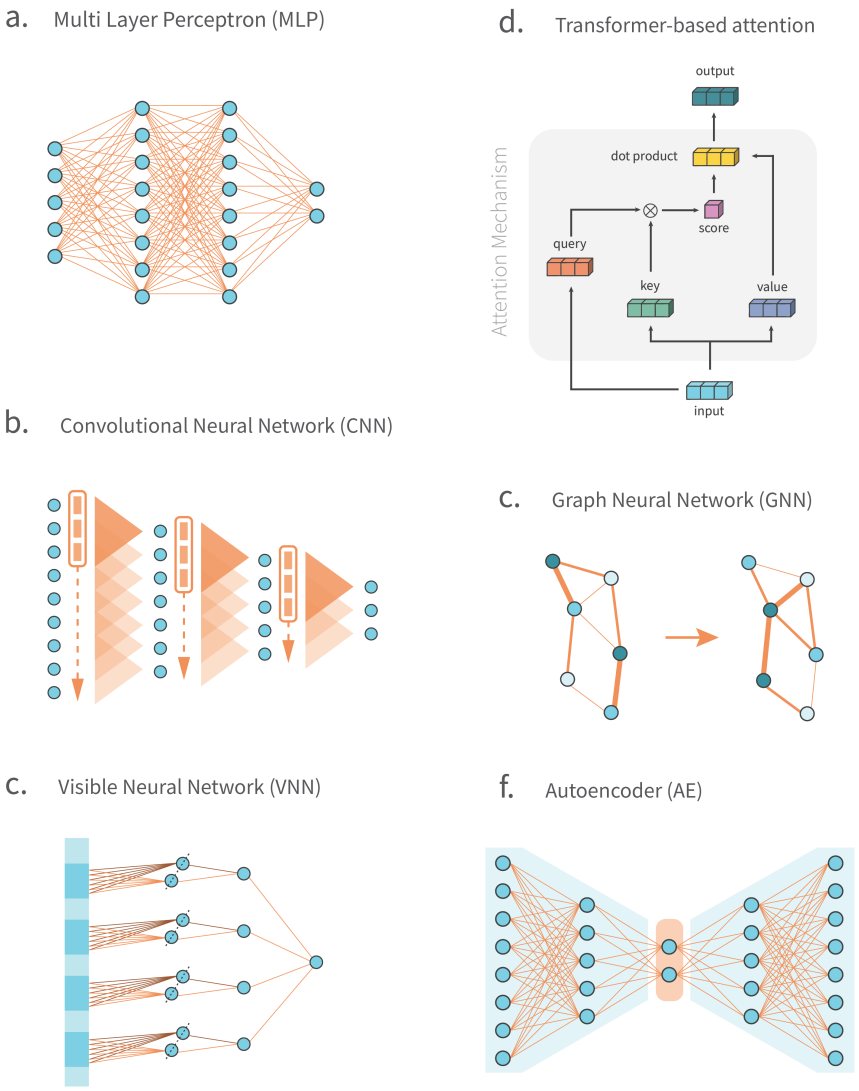


Figure 6.1: **Overview of the most popular neural network architectures. a)** Multilayer perceptron, a fully connected neural network. **b)** Convolutional neural network, **c)** Graph neural network. **d)** GPT; Generative pretrained transformer, here displayed the attention mechanism (inspired by [343]) **e)** Visible neural network and **f)** Autoencoder.

be the length of a DNA sequences or values of reads (e.g. DNase, ATAC-seq data) mapped to a reference sequence.

Graph neural networks (GNN) [347] are designed for analyzing data structured as graphs

(e.g. molecules, proteins social networks, protein interaction networks). A graph neural network is structured like a graph, using prior information to describes which nodes are connected to which nodes (see Figure 6.1c) . For example, each protein in a protein interaction network is a node, with edges between proteins that interact. Message passing, aggregating information from neighbouring nodes, enables each node to aggregate and process information from its immediate neighbors. This process occurs in each layer, where layers can be thought of as iterations of the graph with updated weight. This iterative aggregation enables the network to learn node representations that reflect not only the features of the nodes themselves but also their relationships within the graph. With each successive layer, GNN integrate information from broader neighborhoods, capturing more complex and global patterns in the graph structure.

Visible neural networks (VNN) were introduced in the field of biology to tackle two common problems for deep learning in genomics: dealing in an efficient manner with the large input size and addressing the lack of interpretability. These types of neural networks reduce the number of learnable parameters by embedding biological information in the network architecture so that only biologically meaningful connections are retained (see Figure 6.1d). Each neuron in a visible neural networks represents a biological entity, for example a gene or a pathway [348]. In the network illustrated in Figure 6.1d the highlighted neuron represents a gene, only genetic variants that have a relation to that gene (according to prior biological knowledge) is used as an input. The output of this gene could be connected neurons representing pathways that this gene is involved in. Visible neural networks can be seen as a hybrid between graph neural networks and multi-layer perception. It uses the mechanics of fully-connected feed-forward layers but are shaped like a graph using external sources of biological knowledge.

Generative pretrained transformers (GPT) are the most recently proposed class of neural networks that have had a major impact in research and society [349]. Transformers were developed for the task of translating natural language texts and perform best on sequential data. A transformer alternates between feed-forward layers and the self-attention mechanism. Self-attention (see Figure 6.1e) allows each element in a sequence to dynamically weigh the importance of all other elements. In the original translation task, each word in a sentence can 'attend' to all other words, enabling the model to understand context and relationships within the sequence. In genomics, the equivalents of 'words in a sentence' can be seen as 'genes in a cell' [350]. Generative pretrained transformers are large transformers models that have been trained on massive amounts of data. During training, these models learn to predict the next output in a sequence given the previous inputs. During this self-supervised training procedure, the model gains a deeper understanding of the

data and should, therefore be aware of context. For example, it should be able to infer the gene expression of a masked gene in a cell given the expression of all other genes. These models are also referred to as foundation models, since the pretrained models, with their better understanding of the general concepts, can be used for various downstream tasks with little fine-tuning [351].
Autoencoders (AE) (Figure 6.1f) are unsupervised neural networks consisting of an encoder and a decoder [344]. The encoder maps the high dimensional input data into a lower dimensional latent embedding while the decoder reconstructs the original input from this smaller dimensional embedding. This unsupervised encoder-decoder structure, with such an information bottleneck, allows autoencoders to act as dimensionality reduction tools. Variational autoencoders (VAE) [85] use probabilistic resampling to model the output of the encoder as a distribution over the latent space. An extra regularization term is added in the loss function to encourages the learned distributions to approximate a prior distribution (typically a Gaussian). As a consequence of the probabilistic resampling, each sample can be sampled from a wider area in the latent space as opposed to a single point in regular autoencoders. This results in a coherent latent space that can be used for generating new data points by sampling from this latent space.

6.2.3 Navigating interpretation strategies

We classified the interpretation strategies in each paper by the taxonomy defined by Zhang et al. which utilises three dimensions to categorise approaches [335] (Figure 6.2).
The first dimension divides the interpretation approaches into *active* and *passive* according to whether they require to change the network architecture. Active approaches need a specific configuration of the network to work, for example embedding biological knowledge in the network architecture. Passive approaches do not have this requirement and can be applied post-hoc to (nearly) any network. The second dimension describes the type of explanation that is obtained by using the method. It differentiates between three major types: *logic rules*, *hidden semantics*, *attributions* (ordered by decreasing explanatory power). Methods that extract logic rules approximate the learned function of the neural network by a set of rules. Hidden semantics includes methods that explain the inner state of a neural network. Attribution, the final option for the second dimension, was subdivided further into *gradient-based*, *permutation*, *perturbation*, *game theory*, *attention*, each describing a different mechanism to attach an importance value to each input. The third and final dimension describes the level of interpretability with regard to the input space, differentiating between *global*, *semi-local*, and *local* approaches. Global interpretability refers to understanding the overall decision logic of a model and its behaviour across all

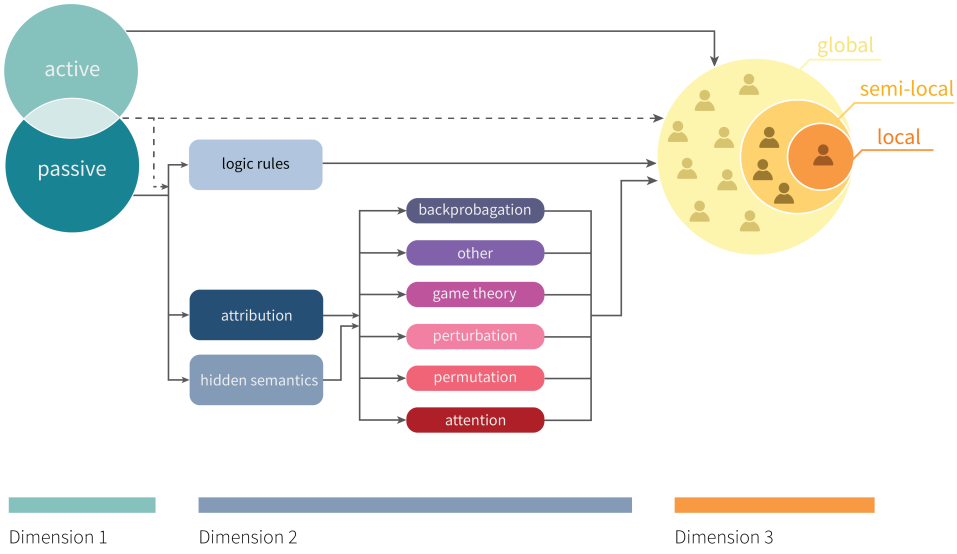


Figure 6.2: **Overview of the three dimensions defined by Zhang et al. [335] to categorise interpretability strategies.** Dimension 1 divides approaches into whether they require to change the network architecture (active) or can be applied post-hoc (passive). Note that active and passive approaches are not mutually exclusive. Dimension 2 further delineates passive interpretability approaches, including methods implementing them. Dimension 3 describes the level of interpretability with regard to the input space, thus differentiating between strategies applicable to whole populations (global), a group of individuals (semi-local), or an individual (local).

examples, for example a global overview of the importance of features over the whole population. Local explanations, provide the interpretation for a single example. To illustate, this could be the contribution of all features to the prediction of a specific patient.[352]. As the transition between global and local interpretation is soft, the category of semi-local approaches describes an intermittent state which can be thought of as e.g. a group of similar patients.

In addition to the taxonomy, we examined which level of biological information was extracted from the network. This can be an inputs (SNPs, genes etc.) or higher level concepts such as gene sets and pathways. Since many of the interpretation methods are novel approaches, we also tagged each interpretation methods based on keywords describing methodological characteristics. Finally, if prior biological knowledge is used in the design of the neural network, we tabulate the source of the prior knowledge.

6.3 The current state of the field: a quantification

During literature retrieval, we employed a systematic literature retrieval and subsequently continuously included newly published articles using the snowballing procedure. In total, our systematic search identified 2008 studies, of which 1146 remained after exclusion of duplicates. During abstract screening, studies were included that i) use human (multi-)omics measurements ii) utilise deep learning architectures iii) attempt to facilitate interpretability. Furthermore, to select the most relevant primary studies for our area of interest, we further excluded studies which used data featuring a spatial (2D/3D images, spatial-omics) or time component (metabolomics, proteomics; except if used in a multi-omics study), as well as studies with a focus on drug development or genome regulatory functions. Additionally, we included relevant articles identified by the snowballing procedure [353], resulting in a total of 123 research articles for our analysis. A detailed description of the literature retrieval procedure can be found in Supplementary Methods. The table with all the surveyed studies is online available as an interactive table for easy navigation, filtering and sorting (`http://www.roshchupkin.org/xai`).

6.3.1 Characteristics of the input data

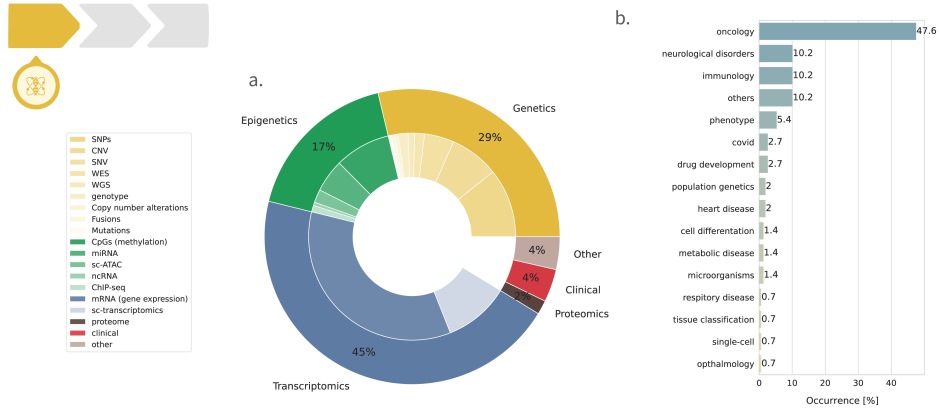


Figure 6.3: **Data types in, and the biological context of, interpretable deep learning studies (121 publications)**. a) Summary on the data types usage. b) Overview of biological fields

Interpretable deep learning solutions have been applied to a wide variety of fields and tasks within genomics. Figure 6.3b shows an overview of fields represented in this study. The majority of interpretable deep learning applications are found in oncology (48%), with neurology (10%) and immunology (10%) following closely behind. Regarding the types of tasks, supervised learning dominates the field, constituting 78% of interpretable neural network applications. Specifically, supervised classification represents 54%, while regression tasks, encompassing survival prediction, make up 24% of the included applications (see Table 6.2). Overall, this demonstrates that the utilization of interpretable models is not limited to specific biological fields or use case but is widely employed across various disciplines.

Figure 6.3a graphically summarises the datatypes found in the surveyed papers. Despite the recent advancement of single-cell sequencing, we found that most studies used bulk sequencing (81%). Single-cell sequencing was mostly restricted to single-cell RNA sequencing ($n = 22$) and to a lesser extend single-cell ATAC-seq (n=4).

Genetic data was used as at least one of the input types for 29% of the studies included in the survey (Figure 6.3a). The small effects and the large number of genetic variants has forced the community to bundle genetic datasets to acquire the sample sizes necessary for these studies. This, together with the relatively low cost of genotyping, results in datasets with large numbers of individuals. The largest sample size among interpretable deep learning applications amounted to 21,105 individuals [354]. However, there is thus a large gap

General					
Publication per year	2023 (13)*	2022 (45)	2021 (32)	2020 (12)	other (19)
Research Field	oncology (70)	immunology (15)	neurology (15)	heritable traits (8)	other (39)
Task	supervised classification (76)	supervised regression (33)	unsupervised clustering (31)		
Data					
Bulk/single-cell	bulk (97)	single-cell (23)			
Data type	transcriptomics (88)	genetics (56)	epigenetics (34)	clinical (7)	other (10)
Sample size	between 500 and 1000 (33)	>1000 & <5000 (29)	>10 000 & <50 000 (19)	>5000 & <10 000 (12)	other (52)
Model					
Model architecture	multilayer perceptron (37)	autoencoder (36)	visible neural networks (24)	graph neural network (17)	other (27)
Number of features	<10000 (40)	<50000 (31)	<1000 (20)	<100 000 (13)	other (16)
Computational resources	unspecified (88)	GPU <13 GB (23)	GPU >13 GB (16)	CPU (11)	other (5)
Interpretability					
Active/passive (1st dimension)	passive (72)	active (35)	passive & active (14)		
Interpertation strategy (2nd dimension)	attribution (101)	hidden semantics (47)	prior knowledge (37)	connection-weights (32)	other (41)
Interpretation methods (2nd dimension)	SHAP (14)	Integrated Gradients (10)	DeepLIFT (9)	Layerwise Relevance Propagation (7)	other (41)
Granularity of interpretation (3rd dimension)	global (61)	local (49)	semi-local (17)		
Level of interpretation	genes (92)	pathways (35)	SNPs (14)	gene sets (11)	other (24)
Source of prior knowledge (active interpretability)	Gene-Onthology (15)	KEGG (13)	Reactome (11)	StringDB (4)	other (32)

Table 6.2: **Overview of the largest categories for each criteria in the main table** (see `http://www.roshchupkin.org/xai` for the full table. Not all categories sum to the total number of papers ($n = 121$) since some studies fall under multiple categories or use multiple methods or datasets. *The systematic literature review was completed before the end of the year

between the sample sizes used in interpretable deep learning applications and the millions of individuals included in GWAS studies (e.g., [355]). Genetic data is sensitive data that cannot be readily shared and gathering large datasets in one place is often infeasible. Distributed learning could be a solution to increase sample sizes [356, 357]. Especially since most common research questions revolve around predicting phenotypes [358–360] and diseases, most commonly cancer [339, 361–366], but also neuro-degenerative diseases [367], psychiatric [368, 369], or hyper-inflammatory [370] conditions. Around these topics, GWAS consortia have formed that have the proper data agreements in place. However, interpretable deep learning with genetic data is not just restricted to these traits and diseases. Examples of other topics for these applications include: differentiating populations [371, 372] or the detecting gene-gene interactions and epistasis [110, 373–375].

Transcriptomics is the most frequently utilised input type covering 45% of the applications. Effect sizes of genes are generally larger than genetic variants and studies can therefore use smaller sample sizes. Sample sizes in the surveyed papers varied between several hundreds [341, 376–378] to tens of thousands individuals [339, 379–382]. Similarly to studies with genetic data, research questions for studies that use bulk transcriptomics data often revolve around predicting various phenotypes based on gene expression differences. Studies that use single-cell sequencing generally have different research questions, and are focused on clustering and integration using autoencoder architectures e.g. [383–386]). Recent popular publications on these tasks use foundation models, which are promising as they can perform several tasks such as cell-type annotation and batch correction. These large generative models are trained on millions of cells, Geneformer [105] used 30 million cells, scGPT [350] was trained with a 33 million cells and scFoundatation [351] on 50 million cells.

Epigenetics is an increasingly popular research area, covering 17% of publication. Epigenomics data are commonly included as one of the inputs in a multi-omics framework - only few publications focus on the sole use of epigenetic data (Supplementary Figure 4). Lemsara et al. [387] combined four omic types in a sparse autoencoder based on pathways and Pan et al. [388] showcased the combination of up to six different data types using a vanilla autoencoder to improve stratification of breast cancer patients. Epigenetic data, often in combination with other data types, have been used to predict cancer states [363, 366, 389–392], drug response [393], COVID-19 [394] or even data types, such as gene expression [377, 395] or even metadata [396].

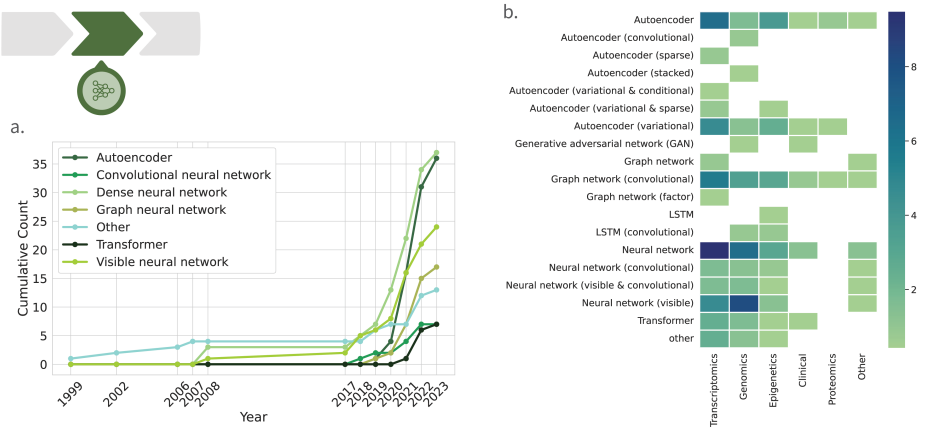


Figure 6.4: **General trends and statistics for model architectures used in the surveyed studies.** a) Number of publications over the years (cumulative) split per neural network type. b) Heatmap illustrating the variety and quantity of data-model combinations.

6.3.2 Neural network architectures underlying interpretable models

Figure 6.4a shows the publication date of these papers categorized per network type. From this figure it is clear that the field is in rapid development, with the vast majority (94%) of papers included in this study being published in 2017 or later. It is also evident that the field has not converged to a single network type, and many types of neural networks are being explored. 6.4b provides an overview of the network architectures per data type. Absence of interpretable deep learning architecture for a certain data type can be a consequence of technical limitations, incompatibility or an opportunity. For example, it will only be a matter of time before interpretable transformer architectures are applied to epigenetics data. At this time, traditional fully connected neural networks (26%) are still the most applied neural network type, closely followed by all the variations of autoencoder (26%), visible neural networks (17%) and graph neural networks (12%).

Fully neural networks can only work with a limited number of input features. When working with expression data this is generally solved by choosing the five-thousand, or less, most differentially expressed genes. However, this loss of data can be avoided by using other network types. Graph neural networks and in particular visible neural networks can learn from larger number of features. Due to the sparely-connected architecture and reduced computational load, visible neural networks are ideal candidates in applications with large number of inputs and limited compute. The largest number of input features,

using more than 4 and 6 million genetic variants, were found in studies employing visible neural networks [369, 397]. Standard graph convolutional neural networks [398], have been used in all data types (see figure 6.4b). Graph neural networks come in a wide-array of variations and can even vary in the way that information from neighbouring nodes are combined. To illustrate, [399] integrated self-attention from transformers in graph neural networks, while [400] used spectral graph convolutions.

All unsupervised learning applications (22.1% of the studies) were clustering tasks, and almost all were applying variations of autoencoders. Here we find notable differences between the use of bulk data and single cell data. For single cell data, there is a data-specific challenge to accurately cluster cell types. Recent articles have proposed to use auto-encoders for this task as autoencoder can provide additional functionalities aside from reducing the dimensionality, for example; remove batch effects, denoise the data, find clusters and integrate multi-omics data [383, 386, 394, 401, 402]. Around 75% surveyed papers using single cell data use variations of autoencoders for clustering.

The concept of visible networks was recently adopted for autoencoders, with the pioneering work of Seninge et al. [386] who designed VEGA, a sparse variational autoencoders for sc-transcriptomics supporting user-defined modules, subsequently inspiring numerous other works. Amongst them was by Lotfollahi et al. [385] who designed the latent dimensions of a sc-transcriptomics autoencoder to represent biological modules with their activities being directly interpretable, further proving the versatility of different autoencoder designs. Using biological knowledge to create more sparse autoencoders, allows these networks to work with more input features with a reduces computational cost, something that autoencoders - with their mirrored design - historically struggled with. It is still an ongoing debate if the encoder, decoder, or both parts of the model should be sparsified [403]. Finally, the latent embedding of autoencoders is known to suffer from the presence of confounding variables [404, 405] and latent features may be entangled, meaning they encode similar information, which hampers direct interpretability.

6.3.3 Interpretability

Dimension 1: Active, passive, or both?

Active interpretability methods require architectural changes, often by integrating biological knowledge, in the network structure prior to training. Naturally, the source of prior biological knowledge is largely determined by the application area of the active network. We found a rich diversity of knowledge sources - around 18 different - during our literature search (see Supplementary Figure 3), demonstrating the big advantage of active

networks, namely to tune the model with respect to ones scientific interest. While around 50% of publications use one of the three major gene / pathway annotation databases (Gene Ontology (GO) knowledgebase, Reactome, or the Kyoto Encyclopedia of Genes and Genomes (KEGG)), others utilise smaller databases describing e.g. miRNA interactions (miRTarBase) [384], functionally associated genes (GeneMania) [381], curated gene sets (MSigDB) [390, 392, 406], or even own data [377, 391, 407]. Active networks can model gene interactions, reactions, or even whole pathways, but in this they are restricted by the quality of the prior knowledge. Integrating incomplete or subjective data may severely limit model performances and interpretability. Active networks are also hampered in their potential to uncover novel biological connections. None of the active deep learning applications had the ability to learn new connections or relations after the prior knowledge was introduced, limiting the viability for niche data types, such as microRNA and non-coding RNA, where only few experimental validated interactions are recorded in databases.

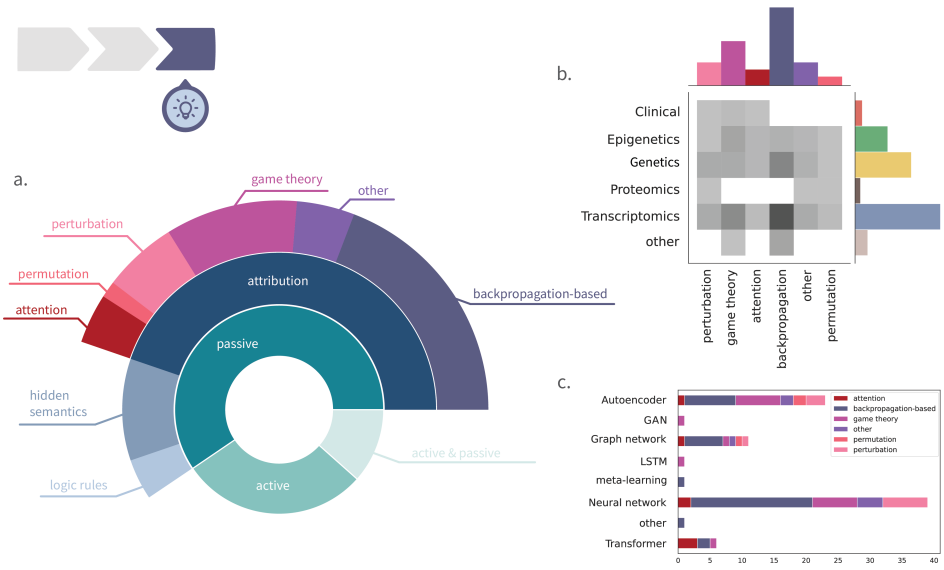


Figure 6.5: **Overview of the interpretability methods and strategies employed**. a) Sunburst plot depicting the strategies employed for the first and second dimensions of interpretability, along with their respective prevalence among 121 surveyed publications. b) Heatmap illustrating the variety and quantity of data-interpretability combinations. . c) Stacked barchart highlighting the identified of model-interpretability combinations.)

Passive interpretability approaches can be applied post-hoc, meaning they do not require researchers to change model architectures prior to training. There are a wide variety

of post-hoc interpretation algorithms, most of them model-agnostic. The latter, allows developers to make easy to use out-of-the-box interpretation solutions that can work for most neural network architectures. Examples of frameworks for passive interpretation are Captum [408], LIME [409], tf-explain [410] and SHAP [121]. The ease of use makes them the interpretabilty method of choice in many studies (passive $n = 72$, active & passive $n = 14$, Figure 6.2a). However, while these methods are flexible regarding the type of model used, most of them provide approximations or make strong assumptions regarding data structures.

Active and passive interpretability approaches are not mutually exclusive. For example, [389] used visible neural networks in combination with DeepLIFT [411], a passive approach. Passive analysis methods can complement active methods such as visible neural networks well. Edge weights do provide global attribution scores but do not provide the important patterns per individual or which features interact.

Dimension 2: How explicit do explanations need to be?

Logic rule sets were found in the earliest applications of interpretability in neural networks [373, 412–414] However, logic rules are not a method of the past; a recent work of Montanez et al. [110] showcased how association rule mining can be used to study epistasis. Association rule mining relies on the construction of sets of SNPs that often occur together across different individuals. By describing the relationship between SNPs in sets, easily understandable logic rules can be derived. Another recent publication, [111], tackled the problem of explainability in healthcare by developing a framework capable of directly extracting rule sets from neural networks, which can subsequently be inspected and adjusted by clinical experts. Deriving rule sets is indisputably the most explicit way of understanding the decision making process of networks. In the case of large or complex networks, however, deriving a reasonable number of (understandable) rules might be infeasible.

Attribution methods, provide less explanatory power as logic rules but are generally quite feasible to apply, as demonstrated by their popularity. Attribution methods, make up 44.8% of the interpretation strategies (Figure 6.2a). As attribution strategies span such a large portion of the interpretation strategies we subdivided them further based on methodological differences. Table 6.3 shows the types of attribution strategies accompanied by a selection of representative application examples.

The largest category, **gradient-based methods**, contains many variations that differ on how the gradient is propagated back to obtain the feature importance. All gradient-based methods surveyed, with the exception of GradCAM, need a reference input to compute

the attribution score. This gives researchers the flexibility to compute importance's with respect to different starting conditions, for example to distinguish between tumour and normal tissue samples, normal tissues can be utilised as reference points [415]. However, the use of inadequate references holds the possibility of spurious or misleading results, as demonstrated in the framework XOmiVAE [416]. In this study, the authors show that their interpretability results vary substantially depending on the reference chosen and conclude that the use of random sets of reference samples is inferior to using normal tissue samples as reference, as the latter specifically highlights important cancer pathways.

Permutation as well as **perturbation** shuffle or change the inputs and observe the change in output. While permutation and perturbation of samples represent a simple way of deriving explanations for any model, they are not well suited for studying large input dimensions or interactions between inputs (e.g. epistasis), as this quickly leads to an exploding number of input combinations to perturb. Perturbations and permutations were used in 11 studies, mainly in for transciptomics and in studies that aim to find epistasis [375]. Sun et al. [417] used local interpretable model-agnostic explanations (LIME) [409] to reveal which genetic variants drive the progression of age-related macular degeneration. LIME utilises perturbed samples to build simple surrogate models approximating the predictions of an underlying black box model, thereby revealing important features [118, 363, 418].

Interpretability methods relying on **game-theory** almost exclusively revolve around Shapley additive exlanations (SHAP) [121]. Its root in cooperative game theory makes SHAP model-agnostic and thus universally applicable - from finding which genes attribute to important pathways in visible neural networks [403], patient-specific feature importance scores for multi-omics cancer data [387], or capturing relevant age-related CpG-CpG interactions with SHAP GradientExplainer, an extension of the integrated gradients method [419]. However, the calculation of SHAP values is complex (NP-hard) [420], so it can be computationally infeasible to deal with high-dimensional data. Additionally, they are designed to handle continuous data and thus show limited support for categorical features, making it challenging to apply them to genetic data.

Attention-based interpretations is mainly used in transformer architectures, but has also been integrated in graph neural networks. Through the use of a graph transformer network and separate attention values for nodes and edges, Kaczmarek et al. [391] were able to determine important miRNAs and mRNAs (nodes) as well as their interactions (edges) in TCGA cancer samples. The use of attention is popular when translating one omics type into another, for example when attempting to predict DNA methylation patterns from genetic sequences, as it reveals relevant regions of interest of the input data type giving

further insight into the interplay of genomic mechanisms [395, 421].
Although the bulk of attribution methods in genomics are adopted from the fields of image recognition or natural language processing, we wanted to highlight some unique approaches stemming from the biological domain. LINA, a linearizing neural network architecture developed by Badre et al. [374] is a backpropagation method capable of delivering first order (individual feature importance), as well as second order (feature interactions) interpretability. Applied to SNP data with the task of epistatis detection, it outperformed other attribution methods such as DeepLIFT or LIME. *DeepResolve* [422] sets the goal of visualising how genetic features interact and contribute to a final phenotype. The method uses gradient ascent and allows negative values in its feature interaction map, thereby addressing limitations of other gradient-based methods.

Table 6.3: **Overview on most widely used attribution methods with a hand-picked selection of manuscripts applying each method**. The full overview table with all entries can be found online (`http://www.roshchupkin.org/xai`) and in Supplementary Materials

Strategy	Methodology	Citation
Gradient-based	integrated gradients	[380, 415, 423, 424]
	DeepLIFT	[375, 383, 389, 407, 425]
	GradCAM	[341, 400]
	Layerwise relevance propagation (LRP)	[366, 426, 427]
	SHAP GradientExplainer	[419, 428]
	DeepSHAP	[429]
permutation	-	[363, 418]
perturbation	LIME	[417]
	modify input	[118, 430]
game theory	SHAP	[387, 403, 431, 432]
attention	-	[391, 395, 421, 433]
other	LINA	[374]
	DeepResolve	[422]
	Diet Networks with element-wise input scaling	[365]

Hidden semantics - should be considered, if the goal is to decipher the inner workings of network rather than focus on what is important for the final result. Exploring which patterns hidden neurons are sensitive to can be easily done when working with active networks. As hidden nodes represent biological entities in sparse networks, such as GO terms [434] or gene modules [385, 386], their activation can be directly interpreted as their activity. In the VEGA framework, authors even propose calculating differential latent variable (gene module) activity, deriving a differential gene expression-like metric. If there is an interest in sequence motives rather than gene sets, Wang et al. showcased that the

examination of convolutional filters in CNN, which act as a 'motif detector,' can uncover known Alzheimer disease-associated patterns [359]. In a versatile framework, Märtens et al. showed the possibility to simultaneously reduce dimensions and enforce clustering in latent space of a VAE [435]. To circumvent the need for additional dimensionality reduction methods like tSNE or PCA, autoencoder with only two latent dimensions were designed, which upon plotting are directly interpretable as x- and y-axis[371, 372]. For studies not directly observing the activation of hidden nodes or employing autoencoders with a minimal latent size, our survey revealed that also post-hoc attribution methods can be used to infer the meaning of hidden nodes. As an example, Janizek et al. utilised integrated gradients in their biologically constrained autoencoder to (i) explain latent feature contribution to reconstruction accuracy (ii) which genes contribute how much to pathways [403]. Also the analysis of single-cell multi-omics data was enabled through the sequential perturbation of latent features in an variational autoencoder, by observing its downstream differences [402].

Dimension 3: from individual explanations to general patterns

Interestingly, during our survey we could not find any preferences in terms of data types, networks, or interpretability methods for inferring local, semi-local, or global explanations - at which granularity interpretability is achieved thus only depends on the goal of each study.

For precision medicine, *local* methods are employed to obtain patient-specific explanations. Popular examples include the investigation of import input features for supervised prediction of a variety of phenotypes, including cancer [339, 380, 397, 436], autism [437], macular degeneration [434], and multiple sclerosis [367], or interpretation of ECG readouts [430]. Besides interrogating the decisions behind predictions for patients, local interpretability can also be used to gain insight in how genes influence the glycophenotype of cells [438], how gene expression and DNA methylation are connected [395], or which genes are best used to approximate the activity of other genes in gene regulatory networks [439].

If not solely the outcome of one individual, but rather a group of individuals is of interest, *semi-local* approaches should be preferred. Semi-local interpretability is of considerable interest when conducting biomarker discovery or survival analysis, as these research questions have the underlying assumption that groups of individuals exist that can be characterised by a unique genomic pattern. Especially in cancer research, namely NSCLC [364, 440], GBM [441], and BRCA [388, 416], we found that by employing semi-local interpretability approaches, the direct characterisation of patient subgroups in terms of

important (multi-omics) features was enabled. In single-cell sequencing, semi-local methods can enable the characterisation of cell clusters or pseudo-bulk aggregates, which is a key point when trying to study tissue heterogeneity [368].
The highest level of interpretation is constituted by *global* approaches. They aim at explaining the network as a whole - these research questions revolve around identifying the most predictive variants, genes and pathways. Linear models, as used in GWAS studies, always provide global interpretations. Neural networks can achieve these interpretations by taking a bottom-up approach of deriving global insight by aggregating all local interpretations when assuming independent and identically distributed random samples [442] or by using global interpretation methods such as inspecting the weights of visible neural network or by extracting a rule set.

6.4 Opportunities and perspectives

There is a plethora of tools and strategies to achieve interpretable deep learning. In this review, we have tabulated and analysed 121 studies of interpretable deep learning applications in genomics. We observe an evolving and growing field, rich with a wide variety of strategies and tools. Overall, the most applied neural network architecture is still the traditional fully connected neural network, closely followed by newer network types such as the autoencoder, visible neural network and graph neural network. Post-hoc interpretation methods, in particular attribution methods, from popular frameworks such as DeepLIFT [411], SHAP [121], and Captum [408] make up the majority of interpretablity approaches. However, with the rising popularity of graph and visible neural networks, the number of applications with active approaches, in which biological knowledge is used to shape the connections in a neural network, will continue to grow.

6.4.1 A lack of diversity and reproducibility within studies

Unfortunately, we found that most studies ($n = 115$) utilise only a single interpretation strategy, only six studies used multiple interpretation strategies [339, 379, 393, 403, 423, 430]. With this wide range of different interpretation methods available, and without a consensus on the best methods, it is worth-wile to apply multiple interpretation strategies to obtain multiple perspectives. As each interpretation methods has its own set of strengths and weaknesses, a combination of interpretation methods will paint a clearer and more consistent picture. Especially the usage of interpretation methods of different categories may complement each other, as global interpretation strategy might miss individual-level or group-level patterns. Local interpretation strategies, on the other hand, may fail to

provide a a clear overview. Even the use of multiple interpretation methods from the same dimension may be beneficial, as some methods are particularly designed to find interactions between features while others are designed to find the most important features. Next to the observation that most studies just apply a single interpretation strategy, we also observe that most studies just apply the interpretation approach once. Neural networks have stochastic elements, and each trained network will inevitably find a different local minimum with different weights. If the goal of a study is to understand the underlying biology, then it is vital to assess the stability of the interpretation. Studies will need to asses if the set of the most important genes or pathways is consistent over multiple runs. In the surveyed papers, we noticed a general lack of reporting regarding the reproducibility, stability, or overlap of results from different passive interpretation strategies. Only a handful of studies have focused on estimating the robustness of their interpretability results [362, 382, 396, 423, 443, 444].

Finally, it is important to consider that neural networks are non-linear, and that non-linear interactions cannot be captured in a single value. Therefore, extracting a set of rules, although harder, provides more value than the popular attribution methods. Extracted rules can provides insight in the number of interactions, the stability of the prediction model, the behaviour of examples that fall outside of the training set. In this regard, the noticeable absence of probabilistic deep learning methods is noteworthy. A well-calibrated certainty estimate could offer a clear indication of whether a method is applied effectively to a sample unlike the training examples.

6.4.2 Future innovations for interpretable deep learning

The majority of the future innovations in interpretability strategies will likely come from adapting established technologies to genomics data. Novel strategies to explain or interpret deep learning will follow from all scientific fields where deep learning is applied. In fields with inherently more interpretable data, such as image data, the validity of the obtained interpretation can be visually assessed. For example, one can overlay an image and the attribution scores and visually asses the plausibility or, for autoencoders, one can generate the resulting images while traversing through the latent space. In genomics, intuitive validation is limited. Simulated data, can bring relief, and the field has provided useful tools for validating new interpretation strategies (e.g., [445–447]). The field of genomics in itself also offers unique opportunities for interpretable deep learning. High quality databases with various types of knowledge (protein-protein interactions, gene and pathways annotations) have been leveraged in various ways to create interpretable neural network architectures. The field has had novel contributions for interpretable deep learn-

ing such as DeepLIFT [411] and many specialized neural networks architectures such as visible neural network architectures [348].

Visible neural networks are promising neural network architectures where all weights are interpretable. Nevertheless, this most likely comes with a cost in performance and in this aspect there is room for improvement. For example, in most implementations genes and pathways are represented with a single neuron. The number of patterns that a network can learn within a gene or pathway is quite restricted. In contrast, convolutional neural networks use commonly between 64 and 512 feature maps. Additionally, the sparsity of the connections limit the number of interactions that these models can capture between genes and pathways. The quality of interpretations is thus strongly dependent on the quality of the biological information embedded. None of the current implementations can compensate well for missing information and here lies an opportunity to balance between a data-driven approach and a knowledge-guided approach. Learning the gap in the prior knowledge may not be easy, but interoperability, for example finding the interacting nodes, can be a tool to aid in identifying the missing connections. Networks that can learn missing connections will not only perform better and provide higher quality interpretation, they will also provide opportunities to fill gaps in biological knowledge.

Generative pretrained transformers (e.g., [105, 350, 351]) will be a popular tool for at least the short-term future. Ease of use, as shown by the popularity of model-agnostic interpretation methods, is a major factor for adoption. Here, autoencoders, graph neural networks and visible neural networks have a disadvantage as they require more expert knowledge in the design phase. Counter-intuitively, transformers are easily adopted, as once trained, they can be widely shared and easy applied. Interpretation for these large complex models is more complex for various reasons. Experiments in the natural language domain have shown that there is often little correlation between important features revealed by gradient-based interpretation methods and attention. Completely different set of attention weights can results in the same prediction [448], and Bastings et al. [449] argues that attention weights reflect the importance of *representations* of inputs rather than the original input themselves and that those representations might already have mixed in information from other inputs. Finally, transformers are often used as an extra preprocessing step that transforms the data before applying an additional network for a downstream task, bringing an extra hurdle for interpretation. Novel interpretation strategies may therefore be required to enable transformer architectures help researchers in understanding the underlying biology in genomics.

In the long-term, we expect more large-scale multi-omics datasets. Integrating multi-omics data is difficult as the data combines the complexities of all the omics types used [450].

Deep learning applications offer unique qualities that are particularly useful for combining omics data. While other machine learning or statistical methods often depend on dimensionality reduction tools, such as PCA, to bring the data to the same dimension, deep learning models can handle multiple inputs of different sizes and use the appropriate layers for each input. Sequence data can be fed through convolutional filters whereas expression data can be processed using attention, or fully connected layers. The hierarchical structure of a neural network - each layer leads to a more abstract representation - provides freedom in when and how to connect separate inputs. Similarly, other types of data such as clinical data, imaging data and patient records can be integrated in a single prediction model [451, 452]. These models will grow is size with the complexity of the data, and the complexity of the task, but distributed learning might offer a solution to acquire the sample sizes necessary to train these large models. Finding novel ways to interpret these large models or to combine this data efficiently in smaller interpretable smaller models, will be the challenge.

6.5 Concluding remarks

There are many ways to bring interpretability in deep learning applications in genomics, and many opportunities to develop novel approaches to interpret and explain neural networks. Aside from the concerns and opportunities raised in the previous sections, we quantified and visualized common solutions and combinations of solutions. We observed an exponentially growing field, rich with a wide diversity of methods and strategies, and we believe that this healthy diversity will inspire the next generation of more interpretable and trustworthy interpretable deep learning applications.

6.6 Supplementary Notes

All Supplementary Material, including figures and tables, are available under: 10.5281/zenodo.11504748.

6.6.1 Data and code availability

All tabulated information can be found online under http://www.roshchupkin.org/xai. All code to reproduce all figures can be found on GitHub: https://github.com/sonjakatz/reviewInterpretability_figures.

6.6.2 Acknowledgements

The authors thank the supporters of this study, namely the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 860895 TranSYS. The authors wish to thank Dr Maarten (M.F.M.) Engel from the Erasmus MC Medical Library for developing and updating the search strategies.

6.6.3 Funding

S.K was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (860895); E. S. acknowledges the funding received 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 personalised medicine—smart combination of pre-clinical and clinical research with data and ICT solutions). G.V.R. is supported by the ZonMw Veni grant (Veni 1936320).

Chapter 7

mEthAE: an Explainable AutoEncoder for methylation data

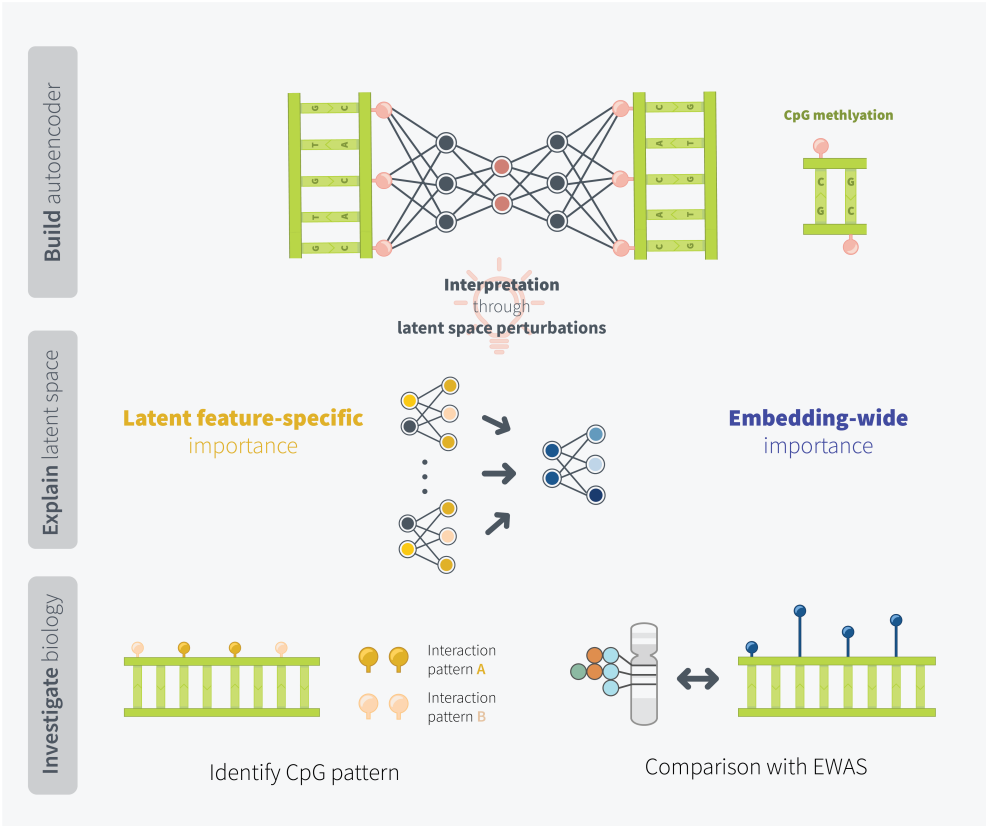
Sonja Katz, Vitor A.P. Martins dos Santos, Edoardo Saccenti, and Gennady V. Roshchupkin

Published in: bioRxiv. 2023 Jul 19:2023-07.
DOI: 10.1101/2023.07.18.549496

Abstract

Background: In the quest to unravel the mysteries of our epigenetic landscape, researchers are continually challenged by the relationships among CpG sites. Traditional approaches are often limited by the immense complexity and high dimensionality of DNA methylation data. To address this problem, deep learning algorithms, such as autoencoders, are increasingly applied to capture the complex patterns and reduce dimensionality into latent space. In this pioneering study, we introduce an innovative chromosome-wise autoencoder, termed mEthAE, specifically designed for the interpretive reduction of methylation data.
Results: mEthAE achieves an impressive 400-fold reduction in data dimensions without compromising on reconstruction accuracy or predictive power in the latent space. In attempt to go beyond mere data compression, we developed a perturbation-based method for interpretation of latent dimensions. Through our approach we identified clusters of CpG sites that exhibit strong connections across all latent dimensions, which we refer to as 'global CpGs'. Remarkably, these global CpGs are more frequently highlighted in epigenome-wide association studies (EWAS), suggesting our method's ability to pinpoint biologically significant CpG sites. Our findings reveal a surprising lack of correlation patterns, or even physical proximity on the chromosome among these connected CpGs. This leads us to propose an intriguing hypothesis: our autoencoder may be detecting complex, long-range, non-linear interaction patterns among CpGs. These patterns, largely uncharacterised in current epigenetic research, hold the potential to shed new light on our understanding of epigenetics.
Conclusion: Our study does not only showcases the power of autoencoders in untangling the complexities of epigenetic data but also opens up new avenues for understanding the hidden connections within CpGs.

Figure 7.1: **Graphical abstract**.



7.1 Background

Epigenetics comprises all changes in gene expression that are independent from DNA sequence alterations. DNA methylation, the most extensively studied epigenetic mechanism [453], involves the addition of a methylation group to a nucleotide, typically the cytosine residue of a cytosine-guanine (CpG) dinucleotide. Epigenome-wide association studies (EWAS) of high-throughput sequencing data have uncovered strong methylation signatures in response to extrinsic factors, such as environmental changes, as well as intrinsic factors such as stress [454–457]. Furthermore, the integrity of the methylome has been closely linked to healthy ageing, as aberrant DNA methylation patterns have been linked to age-related diseases, including Alzheimer disease, cardiovascular diseases, and cancer [458].

The reversible conversion between methylated (hypermethylated) and de-methylated

(hypomethylated) CpG regions has been shown to be highly dynamic [459]. EWAS aim at examining the effect of epigenetic alterations on genome function by identifying DNA methylation variants statistically significant associated with phenotypes of interests. Using general linear models EWAS identify differentially methylated positions by conducting pairwise comparison of mean CpG methylation values between different phenotypes [460]. Despite the wealth of knowledge generated through EWAS [461], such univariate linear analyses are incapable of taking into account the spatial relationships of the up to 450.000 CpG sites measured simultaneously on each array [462]. However, as the genomic distribution of DNA methylation is tightly linked to its gene regulatory functions, gaining a deeper understanding of the relationship between CpG sites is crucial for comprehending how epigenetic modifications influence gene expression and thus ultimately have an impact on human health and diseases [463]. A recent study [464] demonstrated how ageing related CpGs can interact. They show that the contribution of individual CpGs to ageing can be fully dependent on secondary, interacting CpGs. In some instances, a biologically relevant primary CpG site can be completely silenced by a secondary hypermethylated CpG; upon hypomethylation of the respective secondary CpG site, the silencing effect disappears. We therefore argue that it is essential to expand the pool of methods capable of modelling high-dimensional methylomics data, with the goal of finding not only single CpGs of interest, but genome-wide methylation patterns.

One promising approach to overcome the limitations of EWAS in studying the relationship between CpGs is the application of autoencoders to genome-wide DNA methylation data. Autoencoders, and variations thereof, are versatile deep learning frameworks capable of non-linear dimensionality reduction [465–467], clustering [467], data generation and imputation [386, 468, 469], and performing classification and regression tasks [470, 471]. However, the application of autoencoders for whole genome methylation data has proven to be challenging due to their sensitivity to hyperparameter selection [472] and high computational cost [473]. Additionally, autoencoders, as other deep learning frameworks, are inherently not interpretable. Various approaches have been developed to approximate what an autoencoder learns. Most commonly, this involves visualisation of the latent dimension, revealing possible clusters or regions of interest [384, 474, 475]. While autoencoders are frequently being applied to DNA methylation data [465, 466, 470] little work has been conducted on interpreting individual latent features and exploring, for example, which CpGs share a relation through common latent features.

We hypothesise that the way an autoencoder groups together CpGs in its latent dimensions has biological meaning and might reveal novel insights regarding the relationship of CpGs. To explore this, we propose a chromosome-wise autoencoder framework for interpretable

dimensionality reduction of methylation data (mEthAE). In an attempt to validate the performance of our autoencoder, we conduct supervised prediction of age and sex from the latent embedding. Our interpretability pipeline, which is based on latent feature perturbations, yields groups of related CpGs at the latent-feature specific (local), as well as embedding-wide (global) level, which allows us to study CpG relationships at different granularities. In an attempt to put the obtained local and global CpG groups into a biological context, we compare them to EWAS findings, assess their power to predict phentoypes (age), analyse their genomic location, associated biological pathways, correlation patterns, and genomic distances.
Bettering our understanding of the relationship between CpGs could, in the long term, contribute to our understanding of epigenetic mechanisms and have important implications for the development of novel therapeutic strategies for various diseases, as well as for the identification of biomarkers that can help diagnose and predict disease risk.

7.2 Results

7.2.1 Architecture optimisation

To reduce the dimensions of genome-wide DNA methylation data we propose mEthAE, a framework based on chromosome-wise autoencoders. We decided to split our dataset per chromosomes and train 22 individual models, as the high number of input CpGs led to an explosion of trainable parameters if we attempted to train all chromosomes simultaneously. This parameter explosion disabled us from exploring using e.g. multiple hidden layers or larger latent dimensions. With 22 individual models, we were able to thoroughly explore and compare different hyperparameter settings (see Material and Methods). During optimisation, we focused on achieving a good trade-off between size of the embedding and performances, measured as validation loss and reconstruction accuracy which is defined as the Pearson correlation between input and reconstructed CpGs. An overview of the best models, their latent sizes, and reconstruction accuracies for each chromosome can be found in Table 7.1. Overall, we found it possible to compress the input information to a factor of up to 400, while simultaneously achieving reconstruction accuracies close to the maximal possible value (Figure 7.2). We expected the number of optimal latent dimensions to correlate with the size of the input. Interestingly however, although the number of CpGs for each chromosome greatly varies, we observed good performances with latent sizes of 70 across all chromosomes. During optimisation, we often observed a spike in performance around this number of latent dimensions, whereas

performances only increased marginally when significantly more latent features were added (Supplementary Figure 2). To judge whether we encoded biologically meaningful information, we combined the latent dimensions of all autoencoders, totalling 1389 latent features, and conducted supervised prediction of age and sex. Age regression yielded a coefficient of determination of 0.82 and a RMSE of 8.72 years, while the classification AUC ROC was 0.95, clearly displaying the combined autoencoders captured biological signal in their embedding (Supplementary Figure 3).

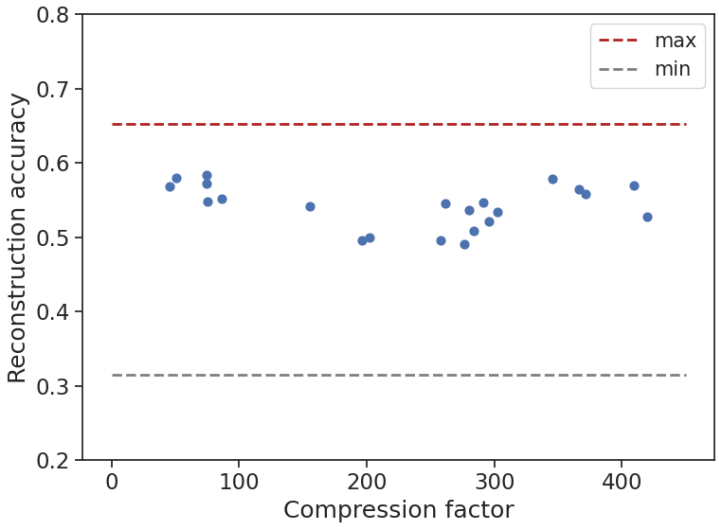


Figure 7.2: **Comparison of reconstruction accuracies in relation to the compression factors for each chromosome.** Red dashed line: reconstruction of models with a latent size equivalent to the number of input CpGs (no compression), averaged over all chromosomes; grey dashed line: reconstruction of models with a latent size of 1 (maximal compression), averaged over all chromosomes. Details on the models with no and maximum compression can be found in Supplementary Table 2.

7.2.2 Embedding-wide (global) CpG connectivity

By assessing the perturbation effects of CpGs across all latent features, we were able to estimate the embedding-wide (global) CpG relationships. We believe that the more impacted CpGs are by perturbations of the latent embedding, the stronger they are connected to (multiple) latent features. This degree of connectivity between latent features and CpGs - in short the CpG connectivity - reflects their importance in the network. CpGs exhibiting similar connectivities form what we refer to as global CpG connectivity groups. We sought to further characterise these global connectivity groups in terms of inter-cluster differences and potential biological relevance.

Table 7.1: **Overview of models after hyperparameter optimisation.**

	Number of CpGs	Latent Size	Reduction Factor	Reconstruction accuracy
chr1	29482	72	409	0.56 ± 0.21
chr2	21984	60	366	0.58 ± 0.21
chr3	15547	45	345	0.58 ± 0.21
chr4	13100	50	262	0.55 ± 0.21
chr5	15965	38	420	0.53 ± 0.22
chr6	23070	78	296	0.52 ± 0.21
chr7	19599	70	280	0.54 ± 0.21
chr8	13704	88	156	0.54 ± 0.22
chr9	6619	88	75	0.55 ± 0.22
chr10	15736	52	303	0.53 ± 0.21
chr11	18570	50	371	0.56 ± 0.21
chr12	15165	52	292	0.55 ± 0.21
chr13	7802	90	87	0.55 ± 0.22
chr14	9417	48	196	0.50 ± 0.23
chr15	9715	48	202	0.50 ± 0.22
chr16	13823	50	276	0.49 ± 0.22
chr17	17044	60	284	0.51 ± 0.23
chr18	3654	72	51	0.58 ± 0.22
chr19	15475	60	258	0.50 ± 0.22
chr20	6567	88	75	0.58 ± 0.21
chr21	2719	60	45	0.57 ± 0.19
chr22	5243	70	75	0.57 ± 0.21
	300000	1389	233 ± 121	0.54 ± 0.21

Number of CpGs: number of input CpGs; Latent Size: number of nodes in latent embedding; Reduction Factor: fold reduction from input size compared to latent size; Reconstruction acuracy: Pearson correlation between input and reconstructed CpGs - indicated for each model by the mean and standard deviation reconstruction accuracy over all CpGs

Degree of global connectivity reflects sample and probe heterogeneity

Our interpretability approach (see material and methods) categorised all input CpGs in three different connectivity groups, shown in Table 7.2. Whereas most CpGs show no or a low connectivity (none, 28.8%; low, 60.9%), a small percentage of CpGs were highly connected across the embedding (high, 10.3%). This ratio of highly connected CpGs was

similar across all chromosomes, indicated by the low standard deviation, whereas the ratio of lowly and not connected CpGs differed substantially between chromosomes. In-depth comparison of these different global groups revealed clear differences between them. When inspecting the range of beta values, CpGs in the highly connected groups showcased a wide range of values, while CpGs in lower global connectivity clusters increasingly converge towards values of 0 or 1 (Figure 7.3A). The inter-sample heterogeneity, indicated through the standard deviation of beta values across samples, correlated directly with the degree of connectivity, with highly connected CpGs showing the highest standard deviations between samples (Figure 7.3B).

Table 7.2: **Characteristics of global CpG groups.**

Global connectivity group	**Ratio** of associated CpGs	**Total number of CpGs** across all chromosomes
global-high	10.3 ± 1.3%	30949
global-low	60.9 ± 14.6%	181798
global-none	28.8 ± 15.4%	87523

Normalised with regard to the total number of CpGs on each chromosome and average across all chromosomes

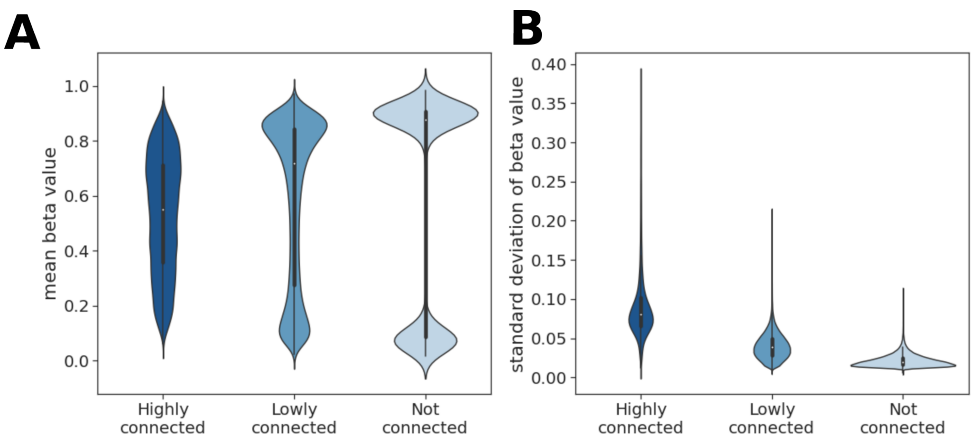


Figure 7.3: **Characterisation of global connectivity groups.** (A) average and (B) standard deviation of CpG beta values across all samples for the different clusters.

Globally highly connected CpGs show higher biological relevance

To estimate whether a distinction in biological relevance between CpGs from different global connectivity groups can be made we (I) interrogated how frequently CpGs from groups are reported in EWAS studies (Fig7.4A) and (II) investigated their predictive power by assessing how accurately participants' age can be predicted using CpGs from the different groups (Figure 7.4B). Comparison to EWAS studies revealed that CpGs from different groups are reported with differing frequencies (Figure 7.4A, dashed lines). To put this finding into relation, we employed a bootstrapping strategy (see Material and methods) (Figure 7.4A, densities). This combined approach revealed that highly connected CpGs are more often reported in EWAS studies than other groups or in the bootstrapped subset. The observed effect is still present, however less pronounced, with lowly connected CpGs and vanishes completely for not connected CpGs. We additionally conducted supervised age regression (Figure 7.4B, dashed lines) and evaluated in a similar bootstrapping approach as outlined above, namely through comparison against the prediction accuracies obtained with randomly drawn CpG subsets of the same size (Figure 7.4B, density). Similar to the picture obtained through the comparison to EWAS findings, it was observed that CpGs with high connectivities are more predictive than other clusters (bootstrap p-value < 0.05), while not connected CpG perform worse than a random subset. As a summary, using two independent approaches, we observed that the global connectivity of CpGs correlates with their biological relevance.

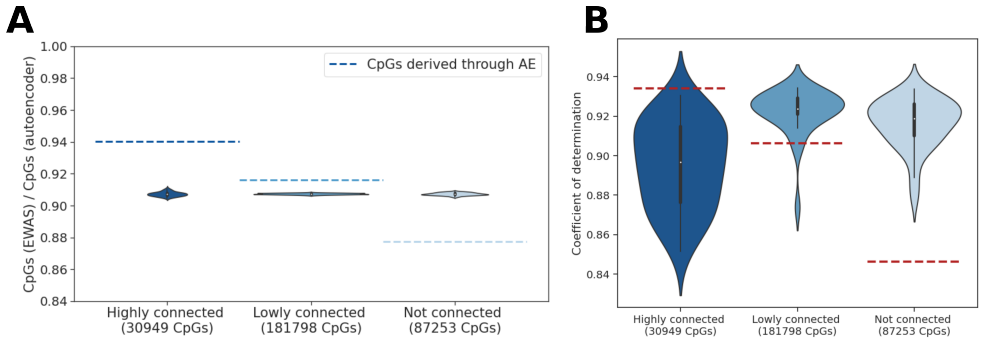


Figure 7.4: **Analysis of biological relevance of globally highly connected CpGs** (A) Proportion of CpGs reported as significant in EWAS catalogue. (B) Supervised age prediction using CpGs from different global connectivity groups. Dashed line - CpGs recovered through interpretation approach; distribution - 100 times randomly sampled CpG set of the same size.

7.2.3 Latent feature-specific (local) CpG groups

In an attempt to gain a deeper understanding of the relationship between CpG sites, we focused on interpreting individual latent features, as we hypothesised that the way an autoencoder groups together CpGs in its latent dimensions has biological implications. By subsequently perturbing latent features and categorising CpGs into different groups based on the observed perturbation effects, we are able to delineate how strongly CpGs are linked to respective latent features. This approach yielded groups of high (local-high), medium (local-medium), low (local-low), and not (local-none) connected CpGs for each latent feature (see material and methods). An overview of the average number of CpGs assigned to each perturbation group can be found in Table 7.3. It is evident that a majority of CpGs are not affected upon perturbing a latent feature (89.4%); only a small percentage of CpGs show high (1.5%), medium (2.2%), or low (6.8%) effects. The e.g. local-high perturbation group only consists of around 208 ± 107 CpG on average per latent feature.

Table 7.3: **Proportion of CpGs in local perturbation groups.**

Perturbation group	**Total number of CpGs**[1]	**Ratio** of associated CpGs[1]
local-high	208 ± 107	1.5 ± 0.2%
local-medium	296 ± 144	2.2 ± 0.5%
local-low	930 ± 475	6.8 ± 1.8%
local-none	12202 ± 6127	89.4 ± 2.2%

[1]averaged across all latent features

averaged across all latent features and chromosomes

Following the observations from the analyses of global CpG groups, namely that globally highly connected CpGs seem to have biological implications, we aimed to investigate this further at a latent feature-specific level, hypothesising that CpGs part of the same local-high perturbation group are functionally or biologically connected. We thus tested whether these CpGs share genomic context associations, biological pathways, correlation patterns, or are spatially located in close vicinity to each other.

Latent feature-specific CpGs are enriched in specific genomic sites

Inspecting the genomic contexts annotated to the local-high CpGs in each latent feature proved that these CpGs are significantly more often located in genomic regions of interest (Table 7.4). Above 40% of latent features and their associated highly perturbed CpGs

can be linked to Enhancer sites or 5‘UTR. Between 10-30% were enriched in DNase I Hypersensitivity Sites (DHS), the first exon, in the vicinity of the TSS (TSS1500), or at the 3‘UTR. It is to be noted that close to no association to CpG island, shelf or shore regions were found. The most common enriched site, with 62%, were gene bodies.

Table 7.4: **Summary of highly perturbed CpGs for each latent feature significantly associated (p-value < 0.05) with different genetic contexts.**

	Latent features associated (p-value <0.05)	**Ratio** [%]
Gene body	861	0.62
Enhancer	622	0.45
5’UTR	586	0.42
DHS	383	0.28
1stExon	305	0.22
TSS1500	299	0.22
3’UTR	223	0.16
TSS200	55	0.04
N-Shelf	26	0.02
S-Shelf	19	0.01
N-Shore	17	0.01
Island	13	0.01
S-Shore	9	0.01

Ratio: Number of features normalised over the total number of latent features (n=1389)

Biological pathways, correlation patterns, or spatial proximity do not explain CpG grouping

Explanations behind the observed grouping of locally highly perturbed CpGs may include their biological roles, a similar co-correlation pattern, or proximity regarding their genetic locations.

To investigate a potential common biological role of CpGs in the local-high perturbation group, we tested for each latent feature if those CpGs are significantly associated with the same Gene Ontology (GO) term or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (FDR < 0.05) (Supplementary Table 3). Out of 1389 latent features, only 201 showed association to (at least one) GO / KEGG term (14.5%). Notably, we observed big differences between chromosomes - while in most chromosomes not more than 15% of latent features could be linked to biological annotations, the CpG groups of every latent

feature in chromosome 6 could be matched to a GO term (100%).
Deep learning models organise information following patterns in the data - we therefore examined possible co-correlation patterns between CpGs belonging to the same perturbation group by calculating their Spearman cross-correlation (Supplementary Figure 4). An average cross-correlation of <0.1 was observed, which suggests the absence of prominent correlations of monotonic associations.
Lastly, we hypothesised that CpGs are strongly encoded in a common latent feature due to spatial proximity on the chromosome, forming linkage disequilibrium (LD)-like clusters. To test this, we examined (I) the median pairwise distances of CpGs as well as (II) defined a LD-cutoff of 0.25 megabases (MB) and calculated a sparsity ratio for each latent feature, whereas a sparsity of close to 1 denote large spatial distances between CpGs (see material and methods). It is evident that CpGs highly perturbed for the same latent feature are spatially not clustered together on the chromosome, with median pairwise distances of >10 MB even for small chromosomes (Supplementary Figure 5A). This observation is confirmed upon inspecting the sparsity ratios for each chromosome (Figure 7.5). Even in the case of increasing the LD-cutoff, only a small decrease in the sparsity ratio is observed (Supplementary Figure 5B). An exception to this is chromosome 6 - it shows low median pairwise distances, a low sparsity ratio, as well as high sensitivity to variation in the LD-cutoff.

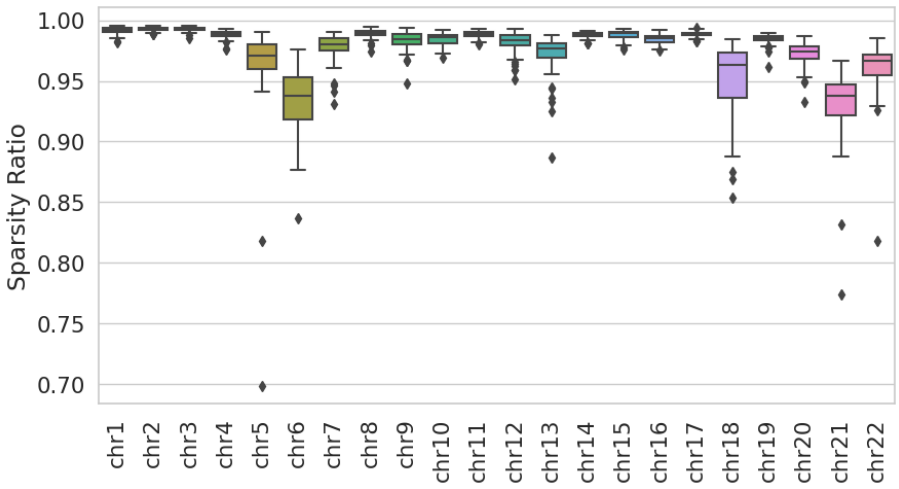


Figure 7.5: **Sparsity ratio of local-high CpG perturbation clusters** for all latent features with a fixed LD-cutoff of 0.25 MB. A sparsity ratio of 1 denotes distances of CpG sites above the designated cutoff, whereas 0 indicates identical positioning of CpGs.

7.3 Discussion

The application of deep learning on DNA methylation data is a promising approach having already made an impact in the field, but it remains challenging to delineate the learned biology from models. We therefore set out to build a framework not only capable of efficiently encoding methylation array data, but also enabling interpretation of how CpGs are grouped together in the latent embedding.

7.3.1 High compression in latent embedding suggests signal sparsity

Although the Illumina 450K array only covers 2% of CpG sites of the human genome [476] and are therefore considered very sparse, our work reveals a high degree of redundancy in the data, demonstrated by the feasibility of compressing the input in a few latent features. We show that just 1389 latent features - as opposed to 300.000 input CpGs, which translates to compression of up to 400-fold - suffice to successfully achieve high reconstruction accuracies. Our results align with previously published deep learning models, which found comparable compression factors [465, 466, 470].

7.3.2 Global CpG groups capture biological diversity

With the aim of interpreting which information is compressed in individual latent features, we developed an interpretability pipeline based on latent feature perturbations. Perturbing or permuting nodes of interest and observing the downstream effects on the network is a widely used attribution method for many data types [335, 477, 478]. A recent publication by Minoura et al. utilised the step-wise perturbation of hidden nodes to find the correlations of genes of interest with individual latent features, ultimately deducing regulatory programs associated with each latent dimension [402]. We took inspiration from these approaches and applied them to methylation array data. To facilitate a more structured interpretation we divided the obtained perturbations effects in two levels: a local, latent feature-specific level, as well as a global embedding-wide level, obtained by grouping CpGs that show perturbations across multiple latent features.

Our analysis of global CpG groups shows a clear pattern: the more these groups are influenced by perturbations of latent features, the stronger their connection to these features, and the more biological information they seem to carry. This becomes evident when we compare groups with varying levels of global connectivity based on their methylation beta values. Firstly, groups with higher connectivity usually contain CpGs with beta values around 0.5, a level often associated with significant variability in methylation patterns

across different cells. Secondly, groups with higher connectivity also display a higher variability of beta values across samples, indicated by a high standard deviation. This is of interest, as typically probes with low variability (low standard deviation) across samples are deemed less biologically significant and are commonly removed during pre-processing. Furthermore, we find that CpGs with higher connectivity are more frequently reported in significant findings from epigenome-wide association studies (EWAS) and tend to show greater predictive power in supervised learning models. Together, this reinforces the idea that (i) the CpG groupings established by our autoencoder is biologically-driven, capturing meaningful relationships among CpGs of different chromosomes and (ii) we are able to observe this with our interpretability method, which suggests that CpGs in the high-global connectivity group may be key in capturing biological variations.

7.3.3 Local groupings suggest potential long range non-linear CpG interactions

Building on the understanding that global CpG groups possess biological significance, our study ventured to explore the reasoning behind the local, latent-feature specific CpG groupings. Our investigations revealed that these CpGs often occupy crucial genomic sites like enhancers, exons, or gene bodies.

This is of particularly interest as it aligns with recent findings indicating the substantial impact CpGs located outside transcription start sites have on gene regulation [479, 480], a departure from earlier focus areas in research [481]. For example, enhanced methylation in gene bodies has been linked to increased gene expression [482, 483], while the methylation of first exons is closely associated with gene silencing [484]. Notably, enhancer methylation has recently emerged as a critical regulatory element, drawing significant attention in genomics [485–487].

Although this discovery resonates with our broader analysis, it doesn't completely clarify the specific relationships within the CpG groupings. Our association of latent feature groupings with biological annotations like GO and KEGG terms indicates that only a fraction can be linked to biological processes or pathways. This suggests the absence of a uniform pattern across these groupings.

Additionally, we found low cross-correlation within these groups together with low ratio of short to mid-range CpG interactions. Instead, we observed that CpGs within the same perturbation group were dispersed along the chromosome, with considerable distances between adjacent sites, suggesting the presence of a long distance regulatory mechanism. A notable exception was chromosome 6, where we observed tighter clustering of CpGs within the same perturbation group, as evidenced by lower sparsity ratios and median pairwise

distances. This aligns with the known genetic density and complexity of chromosome 6, which houses key immune response regulators like the human leukocyte antigen (HLA) complex [488, 489].

While our analyses suggest that the autoencoder groups CpGs based on long-range, non-linear interaction patterns, these patterns are not yet fully characterised in the current research landscape. While strong associations between nearby CpG sites are well known to impact gene expression [463, 490–492] the regulatory role of distal CpG sites is an underrepresented topic because of the difficulty in defining CpG-target gene relationships [493]. However, the importance of long distance CpG sites is not to be underestimated, as it has been shown that most genes are associated with *trans*-methylation sites more than 10 Mb from their promoter region, which explain on average 50% of the variation in gene expression [494]. Long distance CpGs have also been recently proven to be putative biomarkers for subtype-independent breast cancer diagnosis [493]. With the recent finding on the importance of ultra-long-range interactions between regulatory elements [495], we are now prompted to re-imagine the role of long distance methylation marks on gene modulation. However, the challenge to extract novel biological insights from complex data using deep learning remains a daunting task [466].

Our study thus contributes to a deeper understanding of the genomic distribution of DNA methylation and its biological implications, marking a significant step forward in epigenetic research.

7.3.4 Limitations

Due to the computational restrictions we trained individual autoencoder models on each chromosome. Therefore, our current framework can not detect cross-chromosomal interactions of CpGs. While this topic is underrepresented in literature, given our current findings it would be important to test such hypothesis and train one large model for all chromosomes [496, 497].

7.4 Conclusion

In this research, we have successfully developed chromosome-specific autoencoders for methylation array data, achieving a remarkable up to 400-fold reduction in data size without compromising the performance of our models. Our innovative approach, which includes a custom perturbation-based method, allowed us to explore both broad and specific CpG connections within the embedding.

Our finding that we are able to reduce the data size by 400-fold encourages a re-evaluation of thresholds applied in EWAS studies; the redundancies observed in the data suggest that the number of true independent tests may be lower than currently assumed for most studies, which would affect the EWAS significance threshold. A key finding from our interpretability analysis is that CpGs with strong connections across the entire embedding demonstrate significant biological relevance, corroborating findings from epigenome-wide association studies (EWAS). When examining CpGs grouped in individual latent features, we observed that these groupings did not correspond to typical correlation patterns or proximal locations on chromosomes.
This intriguing result points towards the possibility of long-range, non-linear interactions between CpGs, a phenomenon that has not been extensively characterised in epigenomic research. Our study thus opens up new avenues for understanding the complexities of DNA methylation and its role in gene regulation, indicating a rich field of study for future research in epigenomics.

7.5 Methods

7.5.1 Datasets

The data used derived from a continuous ageing study, publicly obtainable from the Gene Expression Omnibus (GEO) under accession number GSE87571. This study by Johansson et al. assessed the methylome of the white blood cells of 732 participants, ranging in age from 14 to 94 years old, using the Illumina Infinium 450K Human Methylation Beadchip [498]. For comparison to the EWAS studies, all metadata and results of the EWAS catalogue [499] were downloaded (version 2022-8-18). For analyses, only studies with results from measuring whole blood samples were included.

7.5.2 Preprocessing

We preprocessed the raw data using the *PyMethylProcess* pipeline, following the steps described in the original *PyMethylProcess* publication [500]. This yielded a total of 300.000 DNA methylation values, called beta values, which are continuous variables between 0 and 1, representing the ratio of methylated versus unmethylated probes for the same locus.

7.5.3 Computational framework

Autoencoder

To reduce the dimensions of the DNA methylation data, our study utilised an autoencoder network, which is an unsupervised neural network consisting of an encoder and a decoder part (see Figure 7.6). While the encoder maps the high dimensional input data into a lower dimensional latent embedding, the decoder attempts to reconstruct the original input from said embedding. We trained the network through a mean squared error (MSE) loss function which quantifies the error between the original input and reconstruction. Prior to training models, the data set was divided into 70% training, 20% testing, and 10% validation.

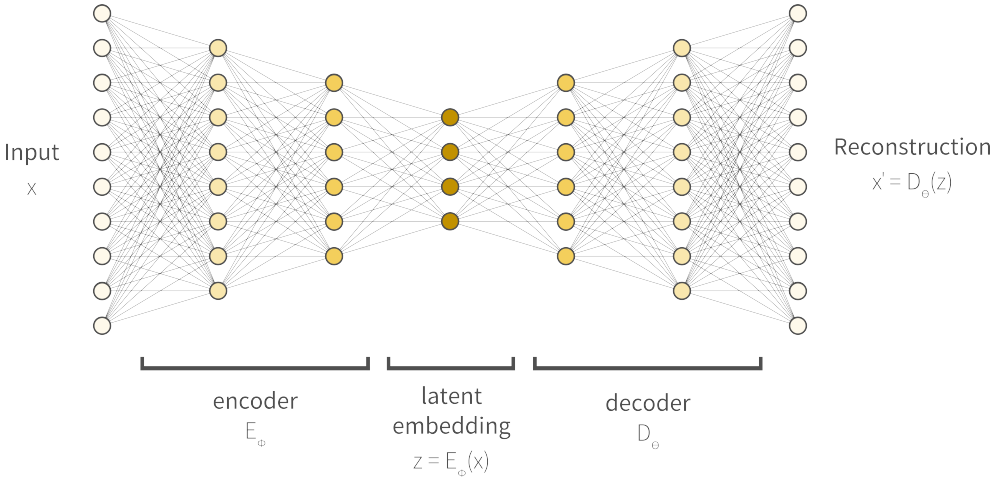


Figure 7.6: **Schematic representation of an autoencoder.** The input (x) denotes a set of input CpGs e.g. 29482 CpGs for chromosome 1. The encoder (E_ϕ, parameterised by ϕ) consists of contiguous hidden layers, each with fewer nodes. The latent embedding (z) represents the bottleneck of the autoencoder with the minimum number of nodes. The decoder (D_θ, parameterised by θ) reversely mirrors the layer structure of the encoder, with the final layer featuring the same number of nodes as the input layer as it attempts to reconstruct the original input from the latent embedding.

Architecture optimization To ensure that we do not compromise model design due to memory issues, we split our dataset chromosome-wise, ultimately training and optimising 22 separate models, one for each chromosome. To derive an optimal architecture for our autoencoders, we carried out an exhaustive hyperparameter search assessing the combination of multiple activation functions, layers (batch normalisation, dropout), learning rates, and number of latent features. All tested architectures were based on densely connected autoencoders with 2 hidden layers before the bottleneck (see Figure 7.6). Between each

hidden layer, reduction factors of 3 and 2 were selected, resulting in a total reduction to 6% of the original input size respectively (Supplementary Table 1). Our hyperparameter optimisation strategy comprised two consecutive parts: firstly, a coarse scan to evaluate a broad range of hyperparameters, followed by a fine scan further optimising the best performing settings.

Coarse hyperparameter scan - During the coarse scan we evaluated latent space sizes with a reduction factor of 2, 4, 8, 16, 32, or 64 respective to the last hidden layer. As a regularisation method to avoid over-fitting, dropout layers with probabilities of 0.1, 0.3, and 0.5 were evaluated. Batch normalisation layers [501] were added to the encoder, which was observed to also enhance training speed and stability. Learning rates evaluated ranged from 0.0001 until 0.005. For the activation function, we used the Parametric Rectified Linear Activation Function (PReLU) [502] in all layers except the output layer, which employed a sigmoid activation function, as beta values are finite continuous values between 0 and 1. Models were trained for a fixed period of 500 epochs using an Adam optimiser [503] and a batch size of 64. As loss function mean squared error (MSE) was used. To evaluate and rank the training performance of models the loss on the validation set was monitored as well as the reconstruction accuracies, measured as the Pearson correlation coefficient (ρ) between the original and reconstructed CpGs. To identify an optimal latent space size from the coarse scan, we evaluated and ranked all combinations of hyperparameters for each latent space size. Subsequently, the latent size showing a satisfactory trade-off between validation loss, reconstruction accuracy, and number of dimensions was selected for further refinement in the fine hyperparameter scan.

Fine hyperparameter scan - To design the latent size grid for the fine hyperparameter scan, six latent sizes in close vicinity of the optimal latent space size from the coarse scan were selected. An overview of the final models including hyperparameters can be found in Supplementary Table 1. Subsequently the fine search followed the same strategy as the coarse search, assessing the identical set of hyperparameter combinations. Ultimately, our optimisation procedure yielded one hyperparameter-tuned autoencoder model per chromosome.

Evaluation of reconstruction accuracies To objectively evaluate whether the reconstruction accuracies (measured by the Pearson correlation (ρ) between the input and reconstructed CpGs) of our optimised models are satisfactory, we wanted to compare them to a version of the model without any compression in the latent space (number of latent dimensions = number of input CpGs), as well a version with maximum compression (number of latent dimensions = 1). In theory, the model without any compression should

achieve the maximum possible reconstruction accuracy, while the model with only a single latent dimension denotes the minimum achievable accuracy.
All models were built in *pytorch* (version 1.10.2) [504] and trained on using the GPU units RTX 2080 Ti (11GB).

Supervised prediction of age and sex

The amount of biological information encoded by the autoencoders was estimated by trying to predict participants' age and sex from the latent embeddings using a supervised model. Therefore, we used the *scikit-learn* [505] implementation of the Random Forest Regressor and Random Forest Classifier. Hyperparameters of the model were left unchanged, except for the number of trees in the forest, which was raised to 1000. Participants with missing annotation were excluded from the supervised prediction. To avoid data leakage, we used the same train-test splits for training the supervised model as were used for training the autoencoder. Scoring for age regression was done by calculating the coefficient of determination (R^2) from actual and predicted age and for sex using the area under the ROC curve (AUC-ROC).

Interpretability of latent features

To understand how the autoencoder encodes input information (CpGs) in the latent embedding, we implemented a post-hoc interpretation approach based on sequential perturbation of latent features which is outlined in the following sections. A graphical summary of the steps is provided in Figure 7.7.
Step (1) - Latent feature perturbations: Latent feature perturbations were carried out sequentially one at a time while holding the remaining latent dimensions fixed. The perturbation value was determined to be one standard deviation of the latent feature activation. Subsequently, other forward passes were conducted, this time with the latent feature activation being altered by the addition or subtraction of the perturbation value. To estimate the effect of the introduced perturbation on each CpG, the median absolute difference between the original, unperturbed reconstruction and the new, perturbed reconstruction was calculated. Finally, the differences over all samples and both perturbation values were averaged, resulting in a single perturbation effect value for each CpG. Repeating this procedure for each latent feature yielded a perturbation matrix with a single perturbation effect value for each CpG and latent feature.
Step (2) - Deriving perturbation effect clusters for each latent feature (local CpG groups): For each latent feature we sought to categorise CpGs into different groups based on the perturbation effects ranging from highly, medium, lowly, to not perturbed groups

- we refer to those as ”local” CpG groups. To categorise CpGs into the aforementioned effect groups, we determined each latent feature's mean perturbation effect and standard deviation and assigned latent-feature specific thresholds (Table 7.5). Subsequently, for each latent feature the perturbation effect of every CpG was examined and CpGs assigned to an effect group using these thresholds. It is important to highlight that CpG assignments were not unique, as these group assignments were performed for each latent feature sequentially.

Table 7.5: **Latent-feature specific thresholds for assignment of CpGs to perturbation effect groups**. Thresholds were derived through determining each latent feature's mean perturbation effect and its standard deviation.

perturbation effect group	**threshold**	
	lower	upper
local-high	$> 3 * std()$	-
local-medium	$> 2 * std()$	$< 3 * std()$
local-low	$> 1 * std()$	$< 2 * std()$
local-none	-	$< 1 * std()$

Step (3) - Deriving embedding-wide perturbation effect groups (global CpG groups): In addition to the latent feature-wise CpG grouping described above, we also sought to infer CpG groups based on all latent feature perturbations combined. We therefore used the effect groups derived from Step (2) to assess the ”global” CpG relationships, which are the perturbation effects of individual CpGs across all latent features. Similar to the thresholding procedure described in Step (2), the mean perturbation effect of all CpG was derived and each CpGs classified into three global relationship groups, termed global CpG connectivitiy groups: globally highly, lowly, and not connected (Table 7.6).

Table 7.6: **Thresholds for assigning CpGs to global connectivity groups**.

global connectivity group	**threshold**	
	lower	upper
global-high	$> 1 * std()$	-
global-low	> 0	$< 1 * std()$
global-none	$= 0$	

As a summary, our interpretability approach yields insights on two levels. Firstly, a detailed insight of which CpGs share a relationship by being linked to common individual latent features. Secondly, an estimate of how strongly CpGs are perturbed across the whole embedding, indicating their overall (global) connectivity in the network.

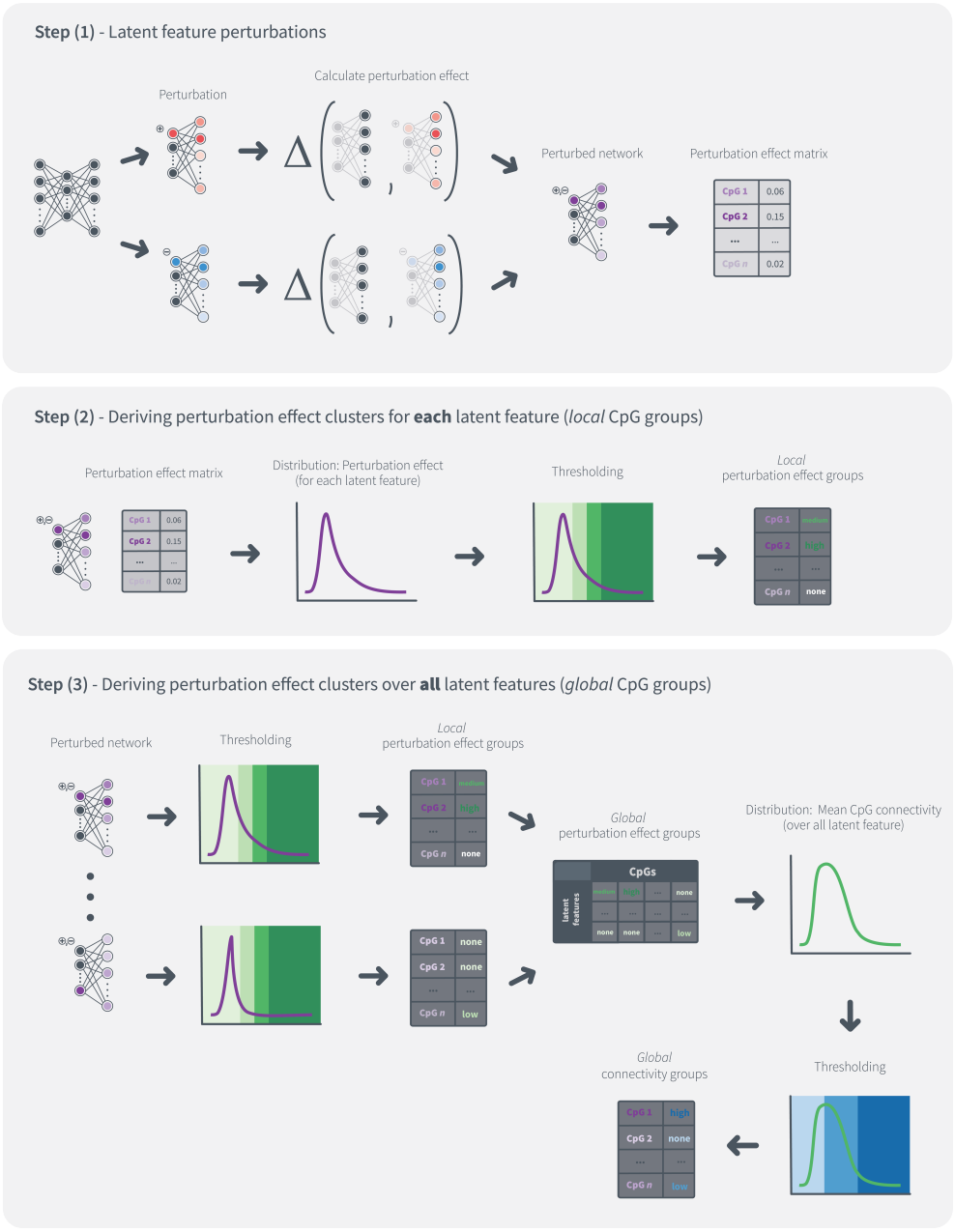


Figure 7.7: **Graphical summary of the interpretability approach utilised in this work which bases on latent features perturbation**. An in-depth description of each step can be found in section 7.5.3. An example of the observed perturbation effects can be found in Supplementary Figure 1.

7.5.4 Characterisation of CpG clusters

Comparison to EWAS findings

A biological validation of the global CpG groups derived by our interpretability pipeline was done through comparison to the EWAS catalogue, calculating how often CpGs from each connectivity group are reported in the EWAS catalogue. To estimate whether the CpGs found by our interpretability method were significantly over- or under-represented with regards to CpGs reported through EWAS, we followed the bootstrapping approach outlined here: (i) all CpGs of the perturbation group under study were checked for significant associations in EWAS studies and the number noted; (ii) a random set of CpGs of the same size as the CpG group under study was drawn and checked for significant associations in EWAS studies; (iii) repeat (ii) for 100 iterations to derive a distribution indicating the chance of a random group of CpGs being mentioned in EWAS; (iv) derive a significance measure by comparing the number of random CpG group configurations that showed the same or higher overlap with EWAS findings than our group of interest.

Investigation of common genomic context, enrichment in biological pathways, correlation, and spatial location

We tested for potential common genetic location or functions of CpGs that are part of the same perturbation group by extracting the genetic information provided in the Illumina Infinium Methylation450K manifest (version 1.2). We applied hypergeometric testing to assess whether CpGs have a common location relative to the nearest gene, association to the same CpG island, location relative to the CpG island, regulatory feature group, DNase I Hypersensitivity Site (DHS), or are located in the same enhancer site. Additionally, using the missMethyl package [506] in R, we tested whether CpGs of the same perturbation groups belong to the same Gene Ontology (GO) terms or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (significance threshold FDR < 0.05). To test for intra group correlation we examined the average Spearman cross-correlation of CpGs in the same perturbation group. To investigate whether CpGs connected to the same perturbation group form LD-like clusters on the chromosome, we assessed their spatial distances by calculating a CpG sparsity ratio. We therefore calculated the pairwise distances of CpGs within the same perturbation group and defined a LD-cutoff of 0.25 MB. The sparsity ratio was subsequently calculated by assessing the number of pairwise distances not within the LD-cutoff divided by the total number of distances.

7.6 Supplementary Notes

All Supplementary Material, including figures and tables, are available under: https://doi.org/10.1101/2023.07.18.549496.

7.6.1 Data and code availability

All code for framework development and performance analysis associated with the current submission is available at https://github.com/sonjakatz/methAE_explainableAE_methylation. Any updates will also be published on Github. Preprocessed data as well as trained models are available upon request.

7.6.2 Acknowledgements

The authors thank the supporters of this study, namely the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 860895 TranSYS. Furthermore we thank the members of the Computational Population Biology group at Erasmus Medical Center for their critical and creative input to this work.

7.6.3 Funding

SK was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860895 TranSYS. GVR was supported by the ZonMw Veni grant (Veni 1936320).

Part 4

Training

Chapter 8

Bridging the Gap in Precision Medicine: TranSYS Training Programme for Next-Generation Scientists

Lara Andreoli[†], Catalina Berca[†], **Sonja Katz**[†], Maryna Korshevniuk[†], Ritchie M. Head, Kristel Van Steen, and the TranSYS consortium

[†]authors contributed equally

Accepted for publication in: Frontiers of Medicine
DOI: 10.3389/fmed.2024.1348148

Abstract

In the evolving healthcare landscape, precision medicine's rise necessitates adaptable doctoral training. The European Union has recognized this and promotes the development of international, training-focused programmes called Innovative Training Networks (ITNs). In this article, we introduce TranSYS, an ITN focused on educating the next generation of precision medicine researchers. In an ambition to go beyond describing the consortium goals, our article explores two key aspects of ITNs: the training and collaboration. Our quantitative analysis approach reveals substantial improvements in scientific, professional, and social skills among young researchers facilitated by the engagement in this interdisciplinary network. We provide case studies underlining the advantages of collaborative environments, featuring innovative scientific exchange within TranSYS. While challenging, ITNs foster positive growth in young researchers, yet exhibit weaknesses such as balancing stakeholder interests and partner commitment. We believe this study may benefit a variety of stakeholders, from prospective ITN creators to industry partners, to design better sustainable training networks going forward.

8.1 Introduction

Precision medicine (PM) represents an evolving field within the healthcare sector, centered on the fundamental premise that an individual's reaction to diseases and therapeutic interventions is intricately shaped by their unique genetic makeup, environment, and a constellation of personalized biological factors. By meeting the specific needs and characteristics of each patient, PM intends to make both clinical decision-making and data communication more effective and to minimize potential side effects [507], and also improve patients compliance. The ambition is to use individuals' phenotypes and genotypes (e.g. molecular profiling, biomarkers, lifestyle data) to tailor the right disease prevention or therapeutic strategies for the right person at the right time. To approach the challenges associated with this paradigm-shift in disease prevention and healthcare, the involvement and collaboration of key stakeholders, including healthcare professionals, academia, policy makers, industry, and patients is required. In order to achieve this goal, understanding the molecular-level causality of pathogenesis becomes crucial. Highly interdisciplinary skill sets are now needed to exploit big datasets and scientific advances that can drive improvements in the prevention, diagnosis, and development of tailor-made interventions for individuals or groups of individuals.

The collaborative nature of precision medicine, which integrates expertise from diverse fields such as genomics, informatics, and clinical practice, demands a commitment to ethical standards. Transparent communication across disciplines guarantees a shared understanding of goals, methodologies, and potential implications, fostering a cohesive and responsible research environment. Ethical considerations extend to acknowledging and addressing biases, promoting inclusivity, and safeguarding against the misuse of data. Socially, a commitment to interdisciplinarity promotes a holistic approach to healthcare that reflects the complex interplay of genetic, environmental, and lifestyle factors. In this way, precision medicine can fulfill its promise of providing tailored, effective, and ethically sound healthcare solutions for individuals and communities.

In this context, Innovative Training Networks (ITNs), and under the EU's Horizon Europe programme the new Doctoral Networks have emerged as valuable frameworks for global research and education collaborations. ITNs are an example of a multinational interdisciplinary PhD programme jointly implemented by academic institutions, industries and others across Europe [508]. These training networks are designed to facilitate transnational cooperation, fostering the exchange of expertise, ideas, and best practices among researchers, and professionals worldwide. Their origin can be traced back to the late 20th century, a period marked by increasing globalization and the rapid evolution of informa-

tion and communication technologies. In response to the need for enhanced international collaboration, the European Commission took a pioneering step in 1990 by launching the Marie Skłodowska-Curie Actions (MSCA) program, which aimed to support mobility and training for researchers within Europe [509, 510]. ITNs revolve around several key focuses, each contributing to the enrichment of research, education and professional development on a global scale [511]:

- Interdisciplinary research
- Mobility and knowledge exchange
- Skill development and training
- Collaborations and consortia building
- Innovation
- Human capital development
- International cooperation

The European Union recognized the importance of developing a skilled workforce that could address current and future challenges and drive innovation in biomedicine. To meet this training need the TranSYS ITN was established to offer Early Stage Researchers (ESRs) the opportunity to gain interdisciplinary training, engage in collaborative research projects, and develop novel techniques and tools with the goal to advance PM [512]. By supporting this interdisciplinary training network, the EU aims to foster innovation, promote scientific excellence, and address critical health-related issues in an increasingly interconnected world. It has evolved from joint workshops and consultations, identifying training gaps in European levels. Preliminary results from implementing pharmacogenomics in clinical settings (cardiology, oncology, and psychiatry) show potential to improve drug use, minimize adverse reactions, and reduce healthcare costs [513, 514].
The consortium at the core of TranSYS consists of a combination of expertise in various fields including Systems Medicine, Functional Genomics, Bioinformatics, Biostatistics, Integrative Biology, Artificial Intelligence (A.I.), Software Development, Blockchain, Ethics, and Pharmacogenomics, completed by industry partners at the forefront of innovative future healthcare. All ESR research projects were complemented by training covering technical (genomics, bioinformatics, health informatics, statistics, data mining, systems medicine and ethics) and key soft skills relevant to high-level research career paths as future leaders for the Precision Medicine revolution. The ESR projects in TranSYS aim to address barriers related to translational systemics (e.g., limited access to validated data

and lack of standards for data storage and curation, along with data privacy concerns) and deliver key innovations to enhance disease detection, diagnosis, treatment, and preventive healthcare.

This paper provides an in-depth exploration of the TranSYS ITN, including its inception, primary areas of focus, and notable achievements. In an ambition to go beyond simply describing goals of the consortium, we aim on giving a detailed insight into two key features of ITNs: the ESR training and collaboration. Using self-reported metrics we evaluate the scientific, professional, and personal growth of ESRs over the duration of the ITN and investigate whether this can be linked to network activities. Furthermore, we measure the scientific output of the consortium and highlight individual projects that could only be facilitated by the collaborative nature of an ITN. Finally, we provide an ESR-perspective on open challenges within ITNs and suggest possible solutions to further improve the experience for early career researchers joining a MSCA-ITN.

8.2 Results

8.2.1 Overview of ESR projects and expertise

The ESRs recruited for TranSYS are as multidisciplinary and diverse as the program itself. Taken together, the ESRs showed a high degree of interaction within the network (Figure 8.1a, chords), but also with the scientific community outside TranSYS (Figure 8.1a , dots). The healthy mix of wet and dry lab projects (Figure 8.1b) enabled the onboarding of ESRs from a broad variety of backgrounds: from physics, engineering, and statistics, to genetics, molecular biology, pharmacology, and ethics (Figure 8.1c). A more in-depth overview of individual projects and how they fit into the frame of TranSYS work packages (WP) can be found in section *Mission and structure of the TranSYS consortium* and Table 8.1.

8.2.2 Professional, scientific, and personal growth of ESRs throughout TranSYS

All ESR individual research projects were complemented by training covering technical (genomics, bioinformatics, health informatics, statistics, data mining, systems medicine and ethics) and key soft skills essential for taking a leading position in the Precision Medicine of the future. With the goal of gaining a more quantitative impression on the achievements of TranSYS in terms of scientific innovation and impact of ESR training, we conducted a survey amongst ESRs. This survey included the questions regarding formal education, a variety of technical, managerial, and soft skills, as well as ratings of official TranSYS

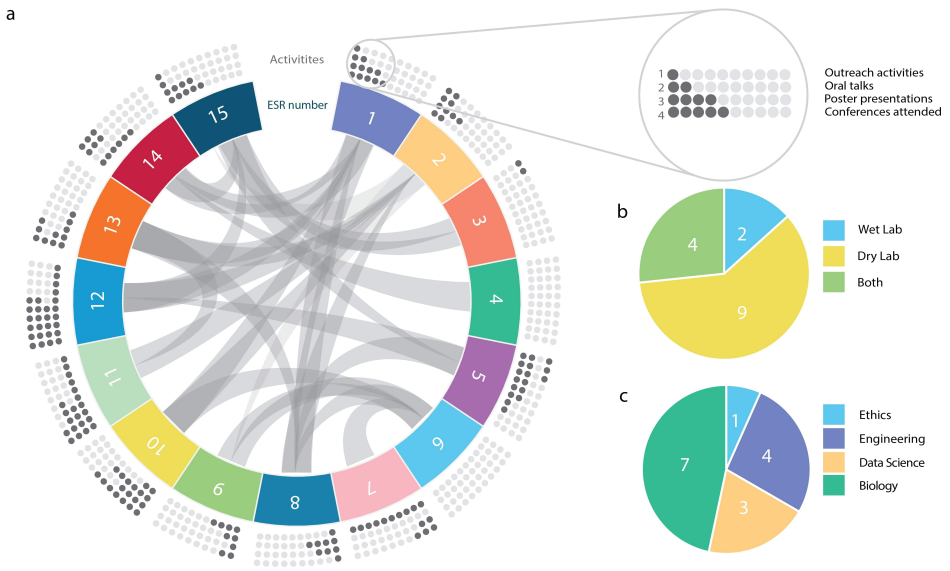


Figure 8.1: **Summary on TranSYS ESRs.** (a) Circos diagram displaying the number (panel), secondments/collaborations (chords), and the scientific output (dots) of each TranSYS ESR. (b) Technical nature of projects. (c) Combined educational background or formal training of ESRs. ESR color coding is consistent throughout the manuscript.

events, such as summer schools and secondments, assessing their perceived impact on the development respective skills.

To assess the development of ESRs throughout the period of TranSYS, we compared the self-reported skill ratings of a set of diverse categories from the start (1st year) to the end of the ITN period (3rd year) (Figure 8.2). A more detailed analysis of each category can be found in Supplementary Figures 1. It is evident that while ESRs strengthened their skills overall, especially their competences in communication (adjusted p-value < 0.05) and research (adjusted p-value< 0.01) significantly improved. Communication skills include public speaking, writing, and presentation skills, while research skills encompass areas such as knowledge of the field, critical reading, or understanding data ownership. Other categories, however, do not reflect this trend, with individuals reporting perceived indifference or even worsening of skills. Foremost, we find a lack of growth in the area of Work-Life-Balance, revolving around topics such as quality of life, healthy time management, or maintaining personal motivation. ESRs also felt like they did not grow a lot in their professionalism, referring to their competence in upholding deadlines, seeking advice, or contributing to a team.

One can argue that during three years spent working in academia an improvement of skills

Table 8.1: **Overview of ESR projects.**

ESR	Affiliation	Title of the project	WP	Lab	Keywords
1	KU Leuven, Belgium	Development of individual-specific molecular networks	2	Dry	gene-based networks, multi-omics
2	KU Leuven, Belgium	Hunting for patient subtypes through image-based phenotypes as biomarkers for major gene effects in medical disorders	2	Dry	genomic data, patient stratification
3	Erasmus MC, Netherlands	GDPR regulation in translational medicine	1,2,3	Both	multi-omics, patient stratification
4	KU Leuven, Belgium	Polygenic Risk Score(s) in the clinic: ethical challenges and stakeholders' perspectives	2	Dry	bioethics, PRS, ELSI
5	University of Ljubljana, Faculty of Medicine, Slovenia	Personalized molecular signatures for modulating progression of metabolism associated liver disease (MAFLD) to hepatocellular carcinoma	1	Both	genome-scale metabolic models, HCC
6	Université du Luxembourg, Luxembourg	Dissecting cellular heterogeneity of Parkinson's disease (PD) related iPS cells during aging by integrated single cell transcriptomics and imaging analysis to identify disease modifiers	1	Wet	PD, scRNA-seq, neurodegeneration
7	Spanish National Cancer Research Centre (CNIO), Spain	Personalized approaches to modulate tumor behavior using vitamin D3	1	Wet	bladder cancer, targeted therapies, patient derived organoid
8	Institut Pasteur, France	Integrated modeling of systemic autoimmune diseases	2	Dry	transcriptomics, inter-individual variability
9	Barcelona Supercomputing Center, Spain	Patient-centric data integration framework for highly dimensional data	2	Dry	data integration, system-level analysis
10	KU Leuven, Belgium	Development of a Federated Blockchain-based Clinical Architecture for Empowering Data Interoperability	1,2,3	Dry	data interoperability, data sharing
11	Max Planck Institute of Psychiatry, Germany	Identification of biological subtypes related to treatment resistant depression	2	Dry	depression, patient subtyping
12	University Medical Centre Groningen, The Netherlands	Multi-omics analysis to delineate drug-response pathways	3	Both	multi-omics, data integration, single-cell analysis
13	Max Planck Institute of Psychiatry, Germany	Understanding stress-responsive molecular networks	3	Dry	psychiatry, PRS
14	The Golden Helix Foundation, United Kingdom	Standardization of disease and population-specific genotyping panel for preemptive pharmacogenomics	3	Both	preemptive pharmacogenomics
15	LifeGlimmer GmbH, Germany	Developing data mining and A.I. tools to better understand patient heterogeneity	3	Dry	A.I., patient stratification

Displayed are the ESR number (ESR), Affiliation, Project title, Associated work package (WP), Technical nature of the project (wet lab, dry lab, both), and Keywords outlining the field of research.

at every level will occur regardless of the possibilities offered within an ITN. To test this, we correlated the growth of each ESR - measured as the difference between skill levels reported in the 3rd and 1st year - with their rating on the perceived value of TranSYS events such as summer schools (Fig 8.3a) and secondments (Fig 8.3b) for each skill category. While research skills seem to have improved independently of summer school activities (Fig 8.3a, left panel), TranSYS events can be clearly correlated with an improvement in communication (right), evident by the large number of ESRs in the green quadrant. Interestingly, although many valued events for their positive impact on Work-Life-Balance, no real improvement of skill was visible (middle); on the contrary, all ESRs reporting a worsening in this category were dissatisfied with the events. What summer schools were lacking in terms of impact on research skills, the secondments clearly compensated for (Fig 8.3b, left panel). A majority of ESRs rated their secondments very positively,

going hand in hand with noticeable learning effects. A similar effect was observed for the categories communication (right), as well as management, and career advancement (Supplementary Figures 2). However, secondments were also perceived to have a negative impact across a majority of ESR's Work-Life-Balance. Especially the subcategories "living" and "effective time management" suffered (Supplementary Figures 2). The data is clear that while secondments are invaluable in improving research skills and an interdisciplinary mindset, they are intense, taking a toll on ESRs by adding stress due to e.g. relocation to another country.

8.2.3 Scientific achievements

Mission and structure of the TranSYS consortium

Although the global market offer promising job opportunities and sustainable career paths, there is a significant demand for skilled researchers who can bridge the interdisciplinary gap between life- and data/computational sciences in the industry. Consequently, the consortium at the core of TranSYS consisted of a combination of expertise in various fields like Systems Medicine, Functional genomics, Bioinformatics, Biostatistics, Integrative Biology, artificial intelligence (A.I.), Software Development, Blockchain, Ethics, and pharmacogenomics, completed by industry partners at the forefront of innovative future healthcare. An eagle-eye view of the structure of TranSYS can be found in Figure 8.4. Projects were outlined and designed by beneficiaries according to their research areas and capabilities, in order to achieve the following three main goals through interdisciplinary collaborations: *A) Narrowing the gap between preclinical performance and treatment benefit* through the integration of preclinical wet-lab observations with in-silico modeling. The aim was to promote the development of more sophisticated disease progression models and significantly enhance discovery and validation of biomarkers. By bridging the gap between experimental and computational approaches, TranSYS ESRs could unlock a deeper understanding of diseases, leading to more effective diagnostic tools and targeted therapeutic interventions. Noteworthy among these efforts are the contributions of *Najjary et al.* (ESR3), who elucidated the omics on immune responses and immune system associated with the development of cancers [515]. Additionally, *Walakira et al.* (ESR5), who applied genome-scale metabolic modeling to integrate transcriptomics data in a human reference model and extract personalized, context specific models for patients with HCC [516, 517]. *Wilson et al.* (ESR6) dissected the cellular heterogeneity of Parkinson's disease related iPS cells during aging by multiomics to identify disease modifiers (manuscript in preparation). Furthermore, *Berca et al.* (ESR7) generated a full-characterized (at the histology, transcrip-

tomic, genetic and epigenetic level) biobank of PDO from bladder tumor samples and use them to study the underlying mechanisms driving resistance to Erdafitinib (manuscript in preparation).

B) Developing integrative strategies and corresponding integrated workflows by taking advantage of top-tier, state-of-the-art data resources across seven disease areas: inflammatory and autoimmune diseases, cancer, non-alcoholic fatty liver disease, neurodegenerative diseases, psychiatric disease, cardiovascular disease, and rare diseases. Cutting-edge data analysis approaches and innovative tools were employed to analyze intricate multi-level datasets, to gain unprecedented insights into these complex medical conditions, driving forward the understanding and treatment of these diseases. The pursuit of our research objectives has led to the development of groundbreaking data analysis methodologies and innovative tools. For instance, *Melograna et al.* (ESR1) explored the potential of interaction, specifically of individual-specific networks or epistasis, to provide personalized insights on an individual's health or disease [518, 519]. *Li et al.* developed a novel multi-view clustering pipeline based on networks and extraneous information [520]. *Andreoli et al.* conducted a systematic review of reasons to identify the ethical and social implications related to the clinical use of polygenic risk scores. They identified a series of normative gaps to be urgently addressed before polygenic scores are implemented in the clinic [521]. *Yousefi et al.* took advantage of network-based machine learning approaches to analyse clinical variation and capturing the dynamics of microbiome data [522–524]. *Mihajlovic et al.* (ESR9) developed novel machine-learning algorithms based on non-negative matrix tri-factorization (NMTF) for integrating and mining time-series single cell transcriptomics data of Parkinson's disease cell line and a corresponding control with molecular networks and bulk omics data. They aimed to uncover novel disease-associated genes, emphasize mechanisms related to disease progression and propose treatment strategies based on drug-repurposing ([525], second manuscript in preparation). The integration method was adapted and used to gain further insights into NRAS mutant melanoma cells' adaptation to treatment [526]. *Lalli* (ESR10) explored the significance of interoperability in model and data integration, spanning multi-omics and multi-modal datasets, and the transformation of raw data into reusable structures. *Taheri et al.* (ESR11) identified ageing biomarkers using network-specific MRI analysis (manuscript in preparation).

C) Improving understanding of patient heterogeneity and developing economically viable patient stratification strategies via the employment of innovative methods and the creation of novel tools to effectively preprocess and analyze the above-mentioned datasets. Through the meticulous examination of population heterogeneity, variations in disease progression and treatment responses could be discerned. These findings served as a foundation to es-

tablish criteria for patient stratification in complex disease areas, enabling more targeted and personalized approaches to treatment. Leveraging advanced techniques, ESRs contribute to a deeper understanding of these challenging medical conditions and pave the way for more effective and tailored healthcare interventions. *Korshevniuk et al.* (ESR12) are dedicated to develop tools for integaring eQTL summary statistics, creating a pipeline for coexpression QTL mapping combining it in a framework for large-scale federated co-eQTL mapping [527]. *Knauer-Arloth, Hryhorzhevska et al.* (ESR13) investigated the impact of genetic variants and glucocorticoid receptor activation on gene expression and DNA methylation and the relationship of these variants with disease risk [528]. *Karamperis et al.* (ESR14) advanced clinical pharmacogenomics in Europe [529, 530] but also deciphered the complex interaction of pharmacogenomic variants linked to adverse drug events across diverse populations, highlighting their substantial influence on commonly prescribed medications [531]. *Katz et al.* (ESR15) focused on the development of artificial intelligence models to capture patient characteristics utilising a variety of data, spanning from biochemical measurements to molecular data [284, 532–534].
While the scientific ideas at the basis of TranSYS are formulated as these three main goals, they are executed in the form of three work packages, namely Preclinical Science and Molecular Medicine (WP1), Systems Analytics (WP2), and Targeted Therapeutics (WP3). For more detailed information on the strategies formulated within each work package can be found in the Supplementary Note 2.

Case studies: how collaboration within TranSYS enabled scientific success

Collaboration, communication, and scientific exchange are values standing at the core of each ITN. TranSYS was no exception to this, and in the following paragraphs we attempt to give a selection of three diverse examples on how collaboration amongst ESRs facilitated projects that would have otherwise been deemed unfeasible; a graphical representation can be found in Figure 8.5. A more complete overview of the scientific output of the consortium can be found in Supplementary Table 1.

Capturing the dynamics of microbial interactions through individual-specific networks *Behnam Yousefi (ESR8), Federico Melograna (ESR1)* [522, 523]
In a perfect example of how the expertise of ESRs can be combined to tackle prior unfeasible projects, ESR1 and ESR8 joined forces to tackle the challenge of personalizing the analysis of temporal microbiome data [522]. The analysis of individual-specific microbiome profiles over time or across conditions is an underrepresented topic in the field, due to the sheer complexity of the data. Combining the scientific vision and programming

skills of ESR8 with the expertise of ESR1 in individual specific networks and statistics, Yousefi and Melograna et al. were able to propose a out-of-the-box way of handling microbiome data, which they refer to as multiplex network differential analysis (MNDA). MNDA is able to resolve microbial dynamic patterns by combining representation learning and individual-specific microbial co-occurrence networks, ultimately uncovering taxon neighborhood dynamics. Supporting the movement of open science, MNDA is publicly available as an R package [523].

A worldwide spectrum of clinically relevant SNPs associated with drug toxicity: Genetic structure and risk characterization using population sequencing data to unravel geographical patterns and spatial trends in prescribed medications *Kariofyllis Karamperis (ESR14), Sonja Katz (ESR15), Federico Melograna (ESR1); manuscript submitted*
The rapidly progressing field of personalized pharmacogenomics presents a significant potential to transform global approaches to patient drug treatments. Under the scientific guidance of a lead researcher (ESR14) and with collaborative efforts from two fellow researchers (ESR1 and ESR15), endeavors were made to identify, categorize, and analyze genetic variations associated with drug-induced toxicity. This initiative laid the foundation for the investigation of the prevalence of these risk-associated genetic alleles on a global scale, offering potential guidelines tailored to specific populations, which could be valuable for regulatory agencies. This collaborative endeavor serves as a compelling example of how the combined expertise of researchers with backgrounds in biology and computational sciences can culminate in a project of substantial significance for both national and international stakeholders. While the biological objectives were formulated by ESR14, the comprehensive analysis of genomic variants was made possible through the data mining proficiency of ESR15. Furthermore, the statistical and programming acumen contributed by ESR1 enhanced the project's communication by generating clear and information-rich visual representations of the findings.

Deconfounding Variational Autoencoders for Multi-omics Data *Zuqi Li (ESR2), Sonja Katz (ESR15)* [533]
The advent of high-throughput sequencing and the analysis of different types of "omics" data has shifted paradigms in biology and medicine. Today the field has developed towards multi-omics approaches, which follows the idea of using ever more data from different sources, aiming to be rewarded with a holistic picture of e.g. the disease under study. However, the integration of multiple data types, especially from different domains, is not trivial and biological meaningful analysis is hindered by another problem well known in epidemiology: confounding. Currently available data integration frameworks are able to

either accommodate multiple data types (from different domains) or correct for biological confounders. ESR2 and ESR15 aim to address this shortcoming by merging these two, currently mutually exclusive, concepts. By combining their expertise in confounder adjustment and clustering (ESR2) and knowledge on developing deep learning models (ESR15), the two ESRs were able to implement a variety of deconfounding deep learning frameworks. This collaboration shows the potential held in matching ESRs strong computational backgrounds and complementary research interests; advanced in-silico projects yielding tools that may benefit the scientific community beyond the boundaries of TranSYS can be realized.

8.3 Discussion

The continuous progression of basic biomedical research, drug discovery and clinical applications, is allowing the implementation of PM in modern healthcare. Today, PM is widely recognized as an integral component of our future healthcare landscape. This realization has brought about a relevant shift in strategy, transforming PM from a scientifically-driven "bottom-up" development that forged its own path, to a "top-down" approach where decision-makers and healthcare systems actively facilitate PM-based approaches. The "bottom-up" approach often relies on regional networks, as exemplified by Sweden and Germany [535]. On the other hand, the "top-down" approach involves national genome initiatives funded by governments, as seen in countries like England, France, and Denmark [536]. This complementary combination of approaches has fostered the integration of PM into mainstream healthcare, ensuring its widespread implementation and impact. A European Partnership for Personalized Medicine was launched in October 2023 to boost research in precision medicine across the European Research Area. This builds on the activities of the International Consortium for Personalized Medicine (ICPerMed) action plan. This partnership will implement the Strategic Research and Innovation Agenda for Personalized Medicine, developed in collaboration with the European Commission. Achieving this ambition and taking a “Systems of Health” approach will require all key players to interact to deliver the societal benefits of PM to patients, citizens and society. At its core this PM strategy needs highly skilled researchers trained in interdisciplinary cross-sector environments, with collaborative cultures and creative mindsets, as pioneered by the TranSYS network and first cohort of researchers [537].
In the past two decades, remarkable strides have been made in the advancement and implementation of PM. During this period, we have observed the rapid emergence of high-throughput genomic technologies and big data analytics, initially employed in research

to uncover the intricacies of disease mechanisms, and later adopted in clinical settings as potent diagnostic tools [507]. Ongoing directions of biomedical research involve analyzing phenomena that occur across multiple scales, from molecular interactions to cellular processes and whole organism behavior. This has also underpinned a deepening understanding of disease and patient heterogeneity and tools to stratify patient populations and cluster diseases with similar aetiologies. Concurrently, novel targeted therapies have been developed, focusing on specific pathomechanisms, particularly in the context of cancer and rare diseases [538]. This progress has paved the way for enhanced diagnostics and personalized treatment, representing a significant change from the traditional "one-size-fits-all" approach to precision healthcare [539]. To tackle these intricate systems effectively, a holistic approach is required, where researchers from diverse disciplines collaborate and pool their expertise.

MSCA Training Networks have established themselves as a valuable part of the Horizon Europe Framework that aims at developing a skilled workforce capable of driving innovation especially in interdisciplinary fields, including PM. The TranSYS ITN was established under the previous Horizon 2020 programme offers young researchers the opportunity to gain multilayered training, engage in collaborative research projects, and develop novel techniques and tools revolving around personalized medicine.

8.3.1 ESRs showed clear improvement in scientific, professional, and social skills relatable to actions within the ITN

To give a more in-depth view on an ITNs most important asset - the researchers - we evaluated how the opportunities within TranSYS impacted the personal development of its 15 ESR. Across the 4 year ITN duration each ESR delivered a core 36 month individual research project and honed their skills in a range of different categories, showing a positive learning trend in 5 out of 6 assessed categories.

Unsurprisingly, amongst the expertise that can be clearly linked to ITN activities, research and communication skills showed the highest degree of improvement. The interdisciplinary nature of an ITN results in the need of clear communication with peers from different scientific fields and sectors. The networks also provide a platform that lends itself to constantly exposing researchers to new knowledge which drives the improvement of research-related skills. Our findings align with a recent evaluation study of the European framework programmes for research and innovation for excellent science, which found that around 80% of ITN fellows consider training options to be good or very good [540]. However, we found that not all skill improvements could be attributed to direct ITN training activities. Some, such as management and leadership skills or professionalism, including skills such

as networking or maintaining positive work relationships, rather evolve naturally in the course of doing a PhD project that is at the core of the ITN. Very notably, the category "Work-Life-Balance" entailing ratings on maintaining a healthy time management, keeping personal motivation or physical and emotional well-being, showed no improvement and could even be negatively associated with the mobility required within an ITN. The implications that can be learned from this finding are further discussed in the Open Challenges within an ITN section of the Discussion.
Although the statistical power of our analysis is limited due to the small number of ESRs, we found that the key training points within an ITN - scientific excellence and advanced communication skill in an increasingly interconnected world - are well reflected in the skill improvement for ESRs.

8.3.2 TranSYS stood out in terms of the large quantity and quality of research outputs

The recent evaluation study of the European framework programmes report a remarkably high quantity and quality of research originating from MSCA actions, and measure an expansion of the professional network due to collaborations within consortia [540]. The scientific achievements of TranSYS reflect that general finding well - at the time of writing this manuscript, a total of 17 different datasets or computational tools have been generated. As of today, the combined number of publications anticipated at the conclusion of the PhD for all ESRs is expected to surpass 50, with the majority of these manuscripts currently in various stages of preparation, submission, or revision. The high number of collaborations outside of the planned secondments paint a picture of extensive collaboration and networking within TranSYS, often initiated by the ESRs. Our selection of *collaboration success stories* clearly illustrate what collaboration within TranSYS looked like, how it was facilitated through the skills of individuals, and which scientific outcomes were delivered. We believe these showcases underline the spirit and opportunities within ITNs, as none of these (and other) projects would have been feasible for individual students, but were facilitated - or even positively reinforced - within TranSYS.

8.3.3 The impact of TranSYS outside academia

EU Framework Programmes, including Horizon 2020, promote several concepts to define societal impact, amongst them the (i) better contribution of research to tackling societal challenges and the (ii) better societal acceptance of science and innovative solutions [510, 540]. We argue that TranSYS ESRs have significantly contributed to both points, with their

focus on tools advancing personalized medicine, with the intention of improved patient well-being. TranSYS featured ESRs solely dedicated to aspects including responsible data storing and model sharing through federated learning and blockchain technology (ESR10) or ethical questions and stakeholders' perspectives on the clinical use of genomic data in clinics (ESR4). With a mindsets towards open and data-driven research and development, all manuscripts by ESRs were published open access, and the underlying data or code is available publicly or upon request. In terms of societal acceptance, ESRs actively participated in a number of outreach and science communication activities. However, ITN outreach activities largely remained on a local and individual level due to the costs and need for having an established name associated with some activities, e.g. international science fairs. To broaden the societal impact of MSCA actions and help bring ITNs out of academia and into society, we feel the need for dedicating parts of the budget towards funding a joint ITN outreach activity. This could be of larger-scale and marketed more effectively than the individual small outreach activities currently organized.

8.3.4 Open challenges within an ITN

In the spirit of adding to the continuous development of the MSCA Doctoral Training Networks, we provide a critical view on current bottlenecks in ITNs, as perceived by ESRs, and give suggestions on how these can be best addressed to further improve the experience of future ESRs.

The act of balancing interests

By its nature, ITNs are consortia with a variety of stakeholders - universities, academic partners, industry beneficiaries and others - each with unique interests that sometimes can not be easily aligned. Despite the unique opportunities offered by this mix of stakeholders, it can result in a demanding and straining program for ESRs struggling to balance university requirements for obtaining a PhD degree, the scientific goals of their supervising PIs, and the competitiveness of industry partners. Secondments with industry partners not hosting ESRs were found to be challenging; a lack of awareness of the project and skills of visiting ESRs, no clear objectives for secondment projects, or the clash of intellectual property interests preventing company knowledge sharing with ESRs who could have contributed positively to further developments in the company. This balancing act is not unique to ITNs within the field of personalized medicine - another biotechnology-focused ITN called YEASTCELL found that the lack of a common ground in industry–academia cooperation may jeopardize the success of entire ESR projects [509, 541]. To avoid conflicts

of interest, they must be identified early in the project planning phase and all stakeholders have to actively communicate their interests to achieve a set of focused and shared ITN objectives. For example, to facilitate the academic requirements of ESRs, doctoral schools enrolling ESRs should acknowledge ITN training with ECTS. Also, industry partners should be given the chance for direct input into the design of the research programme, this secures stronger industry engagement, and a stronger partnership with less potential for conflicts of interest.

Mobility: a challenging gift

A key point making ITNs such exceptional programmes for young researchers are their mobility opportunities. During secondments, ESRs have the possibility to visit other partners from within the network for extended periods, of up to 6 months, to receive training and collaborate in person - an endeavor which is fully funded by the beneficiary institution that employs the ESR. While mobility is the most prestigious part of ITNs and our study can link successful secondments to an improvement in scientific and professional expertise, we also find that they have a negative impact on the quality of life of ESRs (Figure 8.3B). Although aspiring researchers are aware of the mobility requirements within ITNs prior to starting the position, it is easily underestimated the scale of the disruption to life. TranSYS included two secondments within three years that each lasted several weeks and involved relocations. A majority of the difficulties accompanying secondments arise from the challenges associated with relocation to another country, for example extensions of visa allowances or finding short-term housing, or difficulties in continuing ongoing projects next to the new responsibilities within the secondment, especially in the case of wet-lab experiments. We argue that there is a clear need for counteracting the negative trends seen in the "Work-Life-Balance" category with (i) more workshops focused on acquiring skills to improve the quality of life of ESRs and (ii) a stronger sense of responsibility from guest institutions for visiting ESRs. With the elevated rates of depression, burnout, and anxiety in PhD students being well documented [542] useful training included in an ITN may revolve around topics such as effective project management, mindful productivity, stress identification and management, or how to negotiate difficult topics with supervisors. To alleviate some of the stress associated with a secondment, a stronger sense of responsibility or even the formal commitment from hosting institutions to assist ESR is needed in the period leading up to their visit. This is already built into Doctoral Networks planning but based on the experience of ESRs a more proactive approach needs to be encouraged from the project kick-off. A local representative responsible for offering translating skills, communication with officials, or aiding in identifying adequate housing may prove invaluable

for visiting ESRs.

It takes a village - of not just PhD students

The personal perception of TranSYS being very successful in terms of collaboration, scientific output, and shared knowledge aligns well with the numbers identified in the course of this study. However, we argue that the degree of connectivity within TranSYS could have been promoted even more; while ESRs were highly connected through secondments, summer schools, or daily communication channels, the willingness to collaborate lacked amongst some PIs. The COVID-19 pandemic enforced restriction on traveling and in-person events, and while summer schools were mandatory for ESRs, the rules for PIs were less strict, often resulting in few of them joining events. This led to a number of initial project ideas not getting pursued, due to the lack of engagement by supervisors, or changes in the supervisory team leading to conflicts between initial project directions and changing supervisor interests. Collaborations also rarely extended beyond ESRs or PIs to the lab members of beneficiaries; this could be compounded by the individual nature of the PhD projects.
We can conclude that while a demanding programme, the opportunities provided within ITNs are well worth the effort and commitment required to succeed in the dynamic modern research landscape. However, we argue that ITNs are not without faults. Giving attention to some of the points outlined in this section can lead to the design and implementation of better Doctoral Networks in the future. The weaknesses identified can be mitigated without compromising the strengths of MSCA doctoral training programmes, ultimately making ITNs an even more attractive choice for aspiring future leaders.

8.3.5 Future Prospects

Despite the majority of projects coming to a conclusion in 2023, TranSYS is far from ending. Currently most ESRs are in the process of working towards their PhD degrees and looking towards starting new positions in academia and industry all over Europe. However, the multidisciplinary network they built throughout the period of the ITN will not disappear - especially for ESRs seeking to stay in research. The connections and working relationships are certain to contribute towards a successful career start in academia. In the course of the last official consortium meeting, ESRs agreed to the organization of informal yearly reunions to maintain contact in a scientific, but also personal sense. We hope that this way the connections we built through TranSYS will not only stay strong, but even thrive.

8.4 Conclusion

Precision Medicine has undeniably become a cornerstone of our forthcoming healthcare landscape. In this context, MSCA training actions have firmly established themselves within the Horizon Europe Framework, as the driving force of the new generation of highly qualified scientists. Notably, the TranSYS ITN has showcased substantial enhancements in the expertise of its ESRs, spanning research insight and communication proficiencies. The inherent interdisciplinary nature of an ITN necessitates effective communication among peers from diverse scientific backgrounds. This environment also exposes ESRs to a continuous influx of novel knowledge, thereby fostering the refinement of research-related skills. Moreover, the ITN has yielded substantial collaborative opportunities that extend beyond planned secondments. It is crucial to underscore the social impact achieved through outreach and training initiatives, as these actions effectively bridge the gap between ITNs and society at large. Even though some challenges still remain to be overcome, the excellent outcomes derived from this ITN experience across several dimensions and reflected in this work, are progressively laying the foundation for robust networks capable of conducting exceptional basic and translational research for the future.

8.5 Material and Methods

8.5.1 Collection of self-reported evaluation metrics

To assess the development of ESRs during the course of TranSYS, a questionnaire was prepared covering various areas, including background and education, technical skills and competences, academic mobility, collaborations, and scientific output. We additionally utilized the self-assessment survey part of each ESR's Career Development Plan (CDP) - a mandatory tool within all ITNs to track ESR development - to assess the improvement of ESRs in a range of technical and soft skills. This self-assessment survey adapted from the University of Florida allows the evaluation of current strengths and weaknesses. Therefore, we asked ESRs to report their CDP assessment from two different time points: the first CDP filled upon starting the project, and the last CDP filled upon project end, ultimately covering a period of 3 years (with individual starting points). The self-assessment covered the following major categories, each made up by a number of minor categories and rated with scores from 1 (no knowledge) to 5 (expert) (Table 8.2). In an attempt to relate the impact of TranSYS activities (summer schools, bootcamps) on each minor skill category, we asked ESRs to rate the perceived impact of each activity on their development. The ordinal

scale used for this purpose ranged from 1 (negative impact) to 10 (highly beneficial).

8.5.2 Statistical analysis

For statistical analysis of data collected using the self-assessment survey, we aggregated the minor categories to obtain one skill rating per major category for each ESR. P-values for the comparison of 1st and 3rd year ratings were subsequently derived using a Wilcoxon signed-rank test and multiple testing corrected using Bonferroni correction. To investigate whether TranSYS events show an effect on the skill improvement of ESRs, a linear regression for each category was conducted. All analyses were carried out using R (version 4.3.1) and python (version 3.8.5).

Table 8.2: **Overview of major and minor skill categories assessed.**

Research Skills	Work-Life-Balance	Professionalism	Management and Leadership	Communication	Career Advancement
Broad-based knowledge of field	Maintaining openness and curiosity	Identifying and seeking advice	Providing instruction and guidance	Writing for experts in my field	Building transferable skills
Critical reading of literature in the field	Living (physical, emotional, financial)	Upholding commitments and deadlines	Providing constructive feedback	Writing for a lay audience	Identifying career options
Experimental and research design	Effective healthy time management	Maintaining positive relationships	Dealing with conflict	Grant writing skills	Preparing application or valorization materials
Careful record keeping practices	Maintaining personal motivation	Contributing to a team in the office or lab	Serving as a role model	Speaking clearly and effectively	Asking the right questions
Understanding data ownership	Fostering diversity of academic perspectives	Contributing outside the team	Delegating responsibilities in research setting	Using the right body language	Negotiating skills
Demonstrating responsible conduct in human or animal research and publications		Building and maintaining scientific network	Leading and motivating others	Editing own work	Participating in professional service
Identifying research misconduct				Carrying out peer review	Adopting long-term approach to career

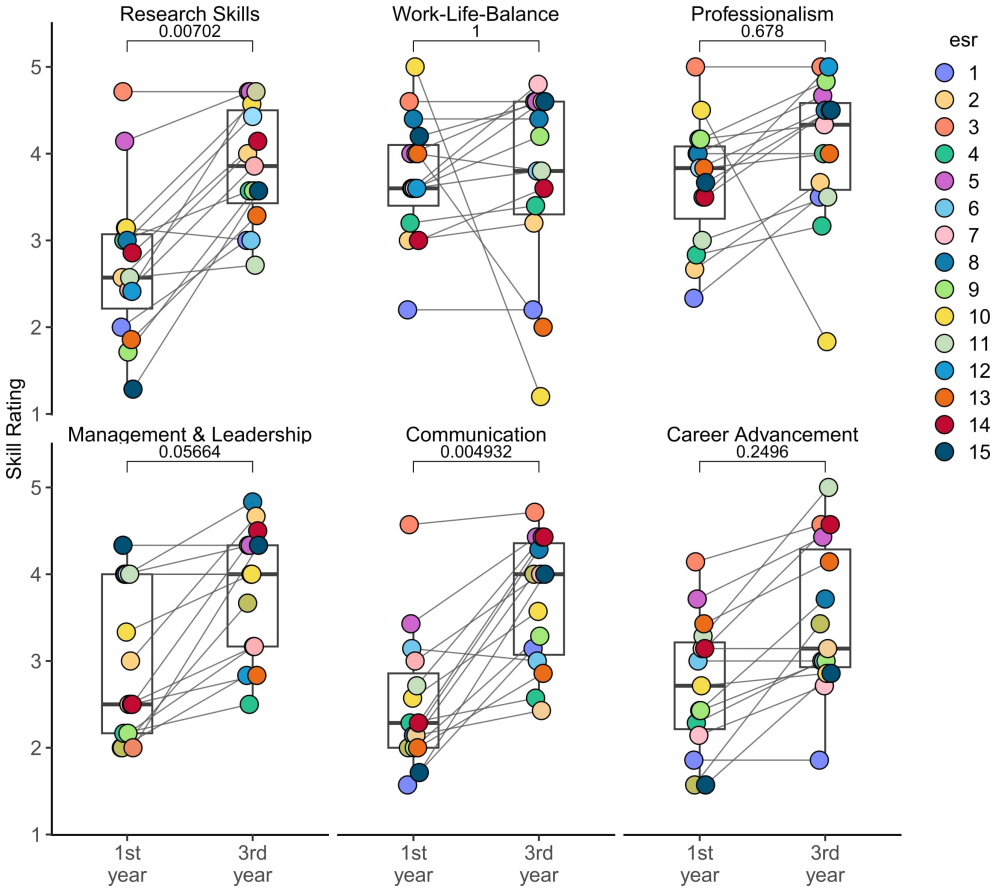


Figure 8.2: **Improvement of ESRs throughout the period of TranSYS in six different major skill categories**. Data was collected using a self-assessment survey upon joining TranSYS (1st year) and prior to rotating off (3rd year) (see Material & Methods). Each dot represents the one ESR, the lines connecting dots represent the relative change in skill rating. Boxplots indicate the summary of all ESRs in respective category and timepoint. P-values were derived using a Wilcoxon signed-rank test and multiple testing corrected using Bonferroni correction. A value < 0.05 can be considered significant.

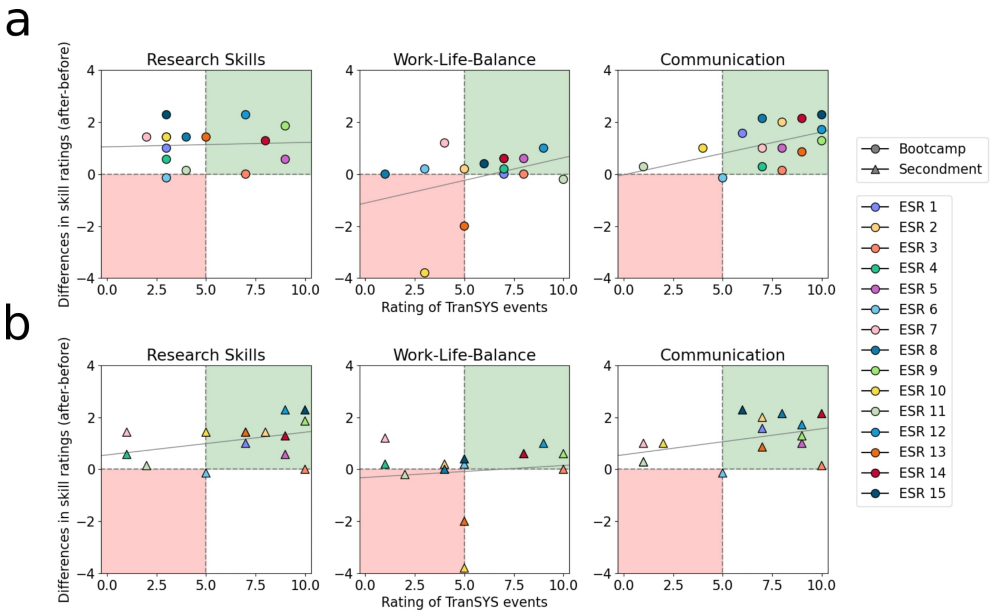


Figure 8.3: **Impact of TranSYS bootcamps (top) and secondments (bottom) on ESR growth**. The x-axis shows ESRs' perceived impact of TranSYS events on the evolution of skills. Skill improvements (y-axis) were derived by calculating the difference in subjective skill ratings of the 3rd and 1st year within TranSYS. Individuals in the upper right quadrant (green) show a positive association between event rating and skill improvement. Individuals in the lower left quadrant (red) display dissatisfaction with TranSYS events paired with a lack of growth in respective skill.

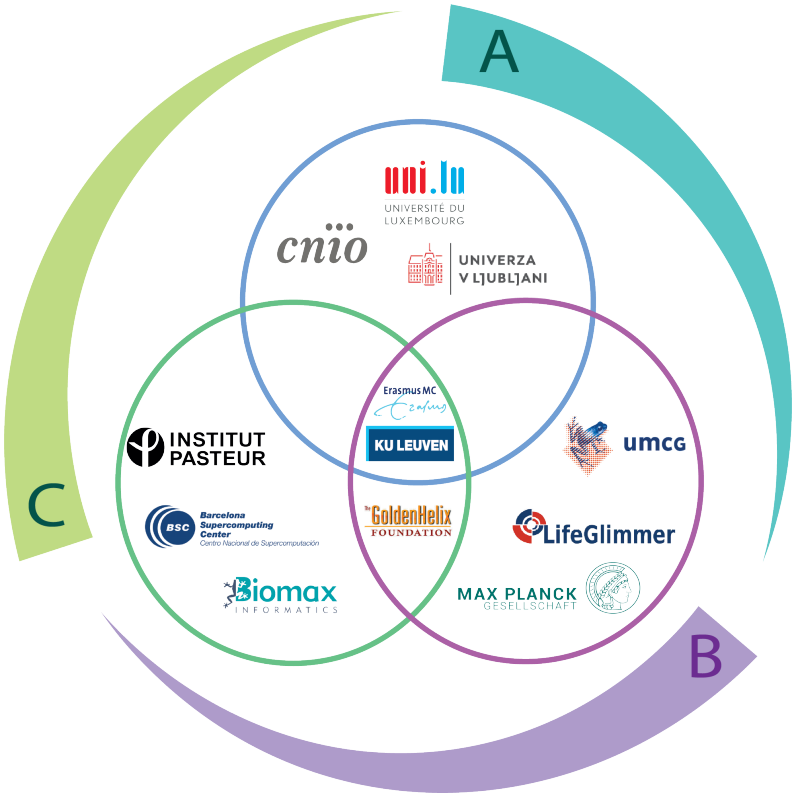


Figure 8.4: **The structure of TranSYS**. The outer circle represents the three key TranSYS objectives: (A) Narrowing the gap between preclinical performance and treatment benefit (B) Developing integrative strategies and corresponding integrated workflows (C) Improving understanding of patient heterogeneity and developing economically viable patient stratification strategies. The inner Venn Diagram illustrates the three work packages of TranSYS and the beneficiaries contributing to each of them: WP1 (blue) *Preclinical Science and Molecular Medicine*, WP2 (green) *Systems Analytics*, and WP3 (purple) *Targeted Therapeutics*.

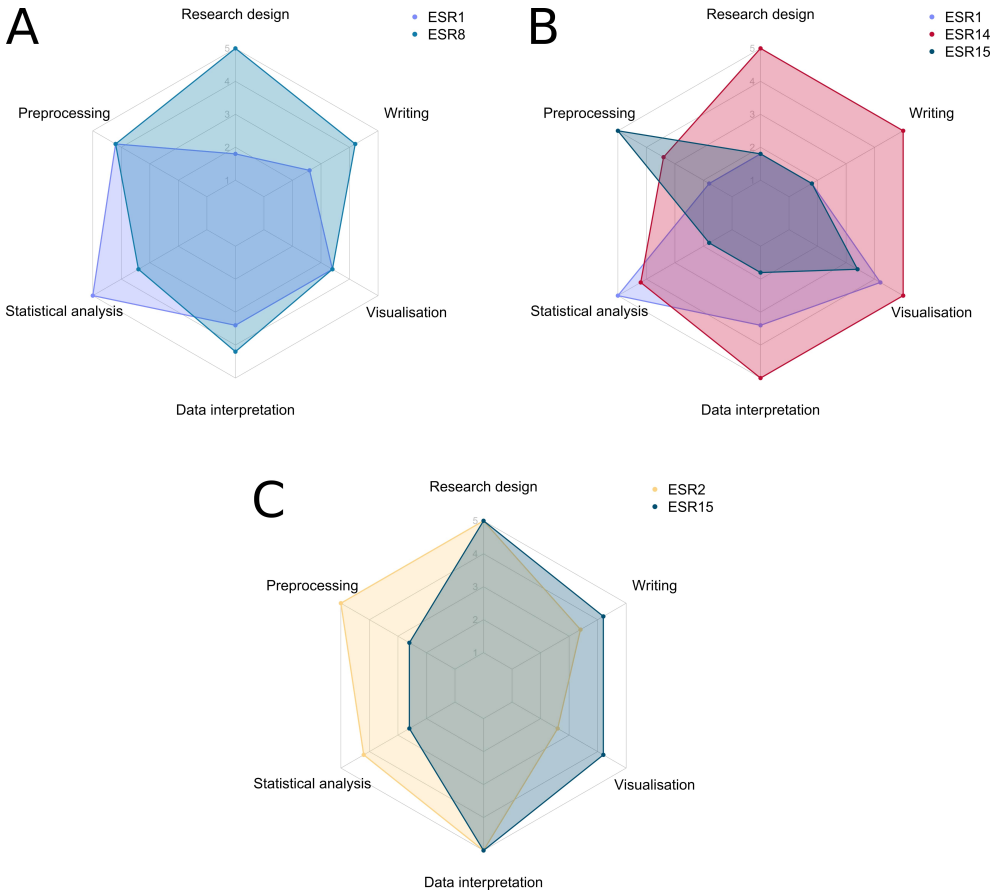


Figure 8.5: **Graphical representation on how ESRs collaborated in successful TranSYS projects.** (a) *Capturing the dynamics of microbial interactions through individual-specific networks*; (b) *A worldwide spectrum of clinically relevant SNPs associated with drug toxicity: Genetic structure and risk characterization using population sequencing data to unravel geographical patterns and spatial trends in prescribed medications*; (c) *Deconfounding Variational Autoencoders for Multi-omics Data*

8.6 Supplementary Notes

8.6.1 Acknowledgements

The authors wish to acknowledge all TranSYS partners who contributed their thoughts and perspectives to this article. We want to especially thank our fellow ESRs for making TranSYS the enriching experience it was. TranSYS ESRs include: Federico Melograna (ESR1), Zuqi Li (ESR2), Shiva Najjary (ESR3), Anneloes Bork (former ESR4), Lara Andreoli (ESR4), Andrew Walakira (ESR5), Elle Wilson (ESR6), Catalina Berca (ESR7), Behnam Yousefi (ESR8), Katarina Mihajlovic (ESR9), Giada Lalli (ESR10), Nahid Taheri (ESR11), Maryna Korshevniuk (ESR12), Anastasiia Hryhorzhevska (ESR13), Kariofyllis Karamperis (ESR14), and Sonja Katz (ESR15). We extend our gratitude to our principal investigators and mentors for their invaluable guidance over the past years, which has been instrumental in fostering the personal growth reflected in this paper: Prof. Dr. Dr. Kristel van Steen (KUL), Prof. Dr. Kris Dierickx (KUL), Prof. Dr. Jan M Kros (EMC), Prof. Dr. Peter van der Spek (EMC), Prof. Dr. Francisco Real (CNIO), Prof. Dr. Damjana Rozman (UL), Dr. Alexander Skupin (ULUX), Dr. Benno Schwikowski (IPASTEUR), Prof. Dr. Nataša Pržulj (ICREA; BSC; UCL), Prof. Dr. Markus Butz-Ostendorg (BIOMAX), Prof. Dr. Lude Franke (RuG), Prof. Dr. Elisabeth Binder (MPI), Prof. Dr. George P Patrinos (UP), Prof. Dr. Vitor Martins dos Santos (LIFEG). We also express our appreciation to Christina Olsen and Ritchie Head from Ceratium BV for their companionship and ongoing support throughout our academic journey.

8.6.2 Funding

TranSYS was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860895.

Chapter 9

Discussion

Parts of this chapter will be prepared for publication

In this thesis, I explored the integration of precision medicine concepts into clinical practice through the utilisation of computational tools (Figure 9.1a). The benefits and potential impact of AI-based frameworks in a variety of topics were demonstrated: from the development of decision-support (**Part 1**) to models enabling unbiased patient stratification (**Part 2**), or showcasing how explainable AI can drive scientific research (**Part 3**). Additionally, training and education programmes for PhD candidates that align with the interdisciplinary nature of precision medicine were presented (**Part 4**). In this chapter, I will highlight how the research done within the frame of this thesis creates value for different stakeholders and thus ultimately contributes to realising the utopic vision that is PM. I will also take a more critical look at precision medicine approaches in practice and discuss, why - 80 years after establishing the concept [543] - the practical use of PM is still limited. Finally, acknowledging that precision medicine's transformative impacts are ongoing, this chapter concludes with a discussion of an upcoming opportunity, which, if seized successfully, has the potential to accelerate the integration of precision medicine into clinical routines and, consequently, our daily lives.

9.1 The Benefit of Precision Medicine: A Matter of Perspectives

Precision medicine encompasses the individual, economical, and societal dimensions. Each facet involves stakeholders whose unique values and needs must align for precision medicine initiatives to thrive. This section aims to provide a contextual framework for the research presented in this thesis, illustrating how our contributions fit into the multi-stakeholder puzzle that is precision medicine (Figure 9.1b).

9.1.1 Accurate and Confident Decision-Making

The most central value for both patients and clinicians lies in the possibility of making accurate decisions with a high degree of confidence [544]. In **Chapters 2, 3, and 4** I present computational frameworks that are designed to aid in the realisation of both these needs simultaneously: clinical decision support systems (CDSS).
When starting the work underlying this thesis in 2020, no tool for the prognosis of NSTI existed; while there had been a diagnostic tool established, its accuracy was subject to debate and with it, its clinical utility in question [145, 545]. Concerning the rarity, rapid progression, and heterogeneity NSTI presents itself with, optimally recognising and handling the disease is challenging, especially for smaller hospitals and non-referral centres.

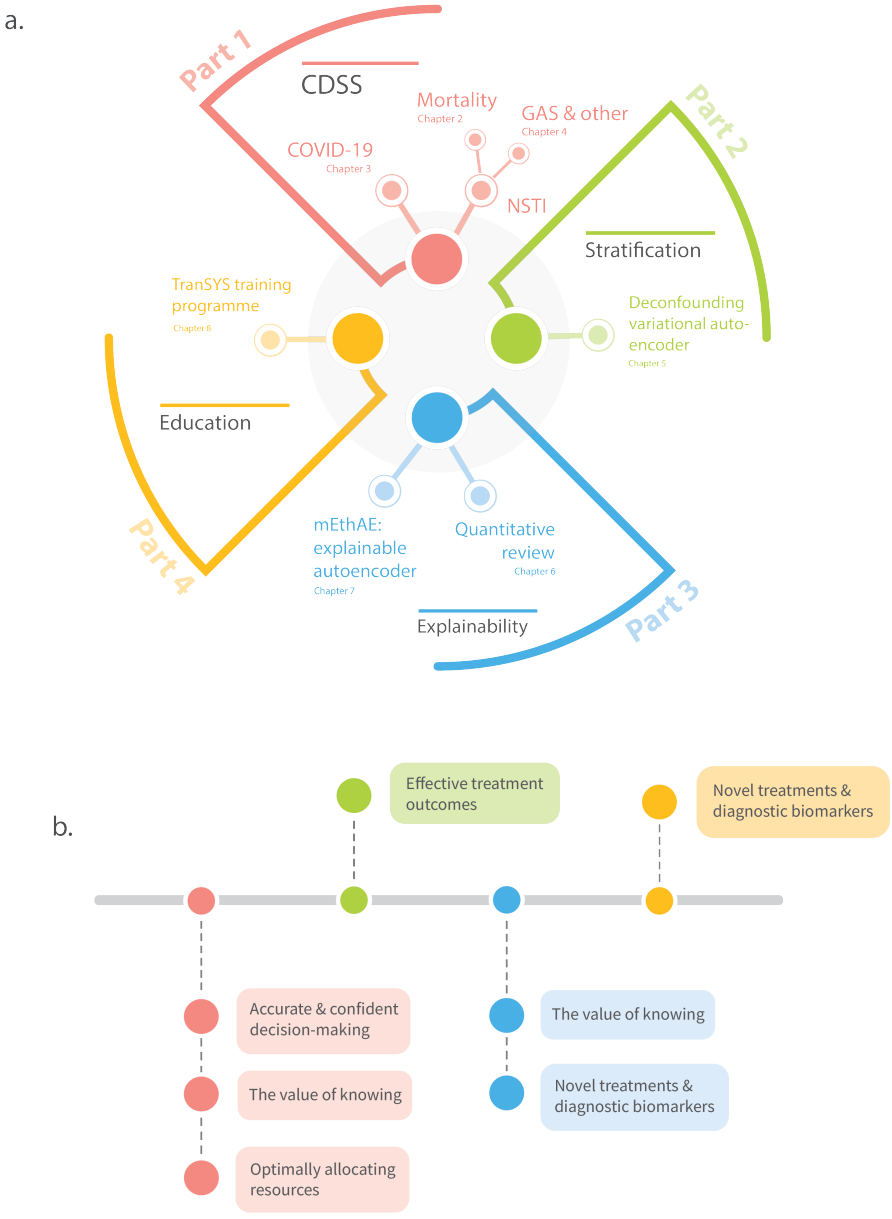


Figure 9.1: (a) **Structure of this thesis**. Each Part of the thesis focuses on a different aspect of precision medicine: Clinical Decision Support Systems (*Part 1*), patient stratification (*Part 2*), model explainability (*Part 3*), and interdisciplinary education in precision medicine (*Part 4*). Omitted are the *Introduction* and general *Discussion* (b) **Summary of the research context and how each part of this thesis relates to precision medicine goals** (section 9.1). CDSS: Clinical Decision Support System, GAS: Group A Streptococcus, NSTI: Necrotising Soft Tissue Infections

Being part of the PerAID/PerMIT consortium, I was fortunate to work in close collaboration with experienced clinicians, which enabled us to develop a CDSS tailored to the needs of NSTI patients and treating doctors. **Chapter 2 and 4** present the fruits of these efforts: not only were we able to provide a mortality prediction tool that uses a basic set of variables readily available even in non-specialised clinics, but we explored a large number of other NSTI-specific clinical outcomes, such as the need for amputation, the anticipated wound size, estimated time spent in ICU, or the presence of specific pathogens, including Group A Streptococcus (GAS).

To the best of our knowledge, our efforts in predicting the presence of GAS represent a forefront in NSTI research which could have significant implications for patients, as this knowledge is crucial in determining early treatment regimens [275].

I was able to translate the knowledge gained from developing a CDSS for NSTI to another infectious disease of high societal importance: COVID-19. In **Chapter 3**, not only did we explore novel biomarkers "outside the box" in the form of sterol intermediates, but we were able to integrate these with conventional clinical parameters into a joint decision-support tool.

Furthermore, our work revealed that CDSS can exhibit limited generalisability across different populations and time - our comparison to COVID-GRAM, a logistic regression model developed in Wuhan, China, showed that while both models use similar parameters to assess disease severity, COVID-GRAM severely under-performs in our Slovenian cohort. This is a well-known problem with CDSS and confirmed by several independently conducted reviews [546–549].

Consequently, I began to question whether the use of external validation is suitable to measure the performance of models that emphasise their adaptability and potential to "learn as they grow". Inspired by researchers sharing this opinion [287, 292], our latest NSTI models presented in **Chapter 4** explored the possibility of providing a localised version of a pre-trained CDSS. By fine-tuning models using hospital-specific data, we effectively re-calibrated them to provide suggestions taking into account the heterogeneity of data across time, geography and facilities, ultimately improving their practical utility.

9.1.2 The Value of Knowing

Increasingly, people seek to understand better what contributes to their personal well-being. Wearable sensors took the consumer market by storm [550] and the global direct-to-consumer genetic test market exploded in the last decade [551]. While an obsession for the data provided by these applications can lead to severe problems such as unintended modification of behaviour or even eating disorders [550, 552], it's worth noting the positive

effects of consumers turning to medical experts, seeking understanding of their results and guidance about managing their health [553]. Clinicians, in turn, need to provide answers to their curious patients, which implies making sense of the often complex and multi-morbid diseases, correctly interpreting analytic results, or justifying clinical decisions.
Therefore, throughout this thesis, all CDSS were equipped with ways to explain the provided suggestions, be it the information on the overall importance of variables through permutations (**Chapter 2**), or delineating variable importance for individual patients through game-theoretic approaches (**Chapter 3 and 4**).
By prioritising the use of explainable AI models in our CDSS, we aim to foster trust among clinicians considering the use of our tools. Additionally, this empowers patients by providing them with access to crucial insights about their well-being, allowing them to derive personal benefits from the wealth of data collected. With the growing use of deep learning models in decision-making, interpreting them has become increasingly important.
Chapters 6 and 7 delve into methodologies that attempt to delineate the decision-making processes within more complex model architectures.

9.1.3 Effective Treatment Outcomes

Realistically, the large heterogeneity in which diseases manifest themselves makes it nearly impossible to provide the ideal treatment for every patient at all times. This translates into a constant act of balancing the trade-off between under- and over-treatment of patients is happening. For instance, the guidelines for NSTI care emphasise that the cornerstone of managing the disease is immediate, aggressive, and radical surgical debridement [554]. This underscores the implication that under-treatment or delayed interventions can significantly increase the risk of fatality for patients. However, aggressive surgical interventions, often including amputations, come at a significant loss in the quality of life of patients [555, 556]. Additionally, from an economical standpoint, over-treatment in hospitals has been reported to account for up to 30% of health care costs [557], translating to billions of dollars.
These examples should illustrate that it is not only in the interest of clinicians and patients but also healthcare providers to determine with the highest accuracy possible, which group of patients will benefit most from a certain treatment.
To deliver a realistic estimate this stratification ideally takes into account a range of diverse patient data such as demographic, clinical, multi-omics, and longitudinal data [558–560]. To integrate and interpret this diverse set of information, modern patient stratification tools commonly utilise AI.
However, it is acknowledged that the quality of stratification, especially across diverse

patient populations distinguished by factors such as age, gender, or ethnicity, is highly dependent on which data models were trained on [561].
To provide truly accurate patient stratification, and therefore maximise the effective treatment outcomes in our increasingly diverse societal landscape, the model should be free of so-called confounders, which are external factors unrelated to the condition of interest. While no model is ever truly devoid of confounders [562], known biases can be corrected during model development.
Working towards the ambitious goal of an unbiased patient stratification tool operating on heterogeneous and multi-dimensional data, in **Chapter 5**, we thoroughly explored and compared different patient stratification models capable of removing confounders in multi-omics data. While previous studies have addressed patient stratification using multi-omics data [563] and the treatment of confounders separately from each other [299, 564, 565], we believe these concepts can not - and should not - be addressed independently.
While our models exhibited major differences in performance, illustrating the complexity of the topic, we were able to develop a framework capable of accurately recovering true patient labels in the presence of confounders while conserving meaningful biological associations. With the trend for Big Data Analytics and even bigger models dominating the current technological landscape, I believe our research addresses a topic that will only continue to gain importance in the future to come.

9.1.4 Optimally Allocating Resources

This thesis focused on two acutely infectious diseases, namely NSTI and COVID-19, which both require intensive care unit (ICU) resources due to the need for close monitoring. The critical infrastructure in hospitals, such as the number of ICU beds, can be limited; especially during the COVID-19 crisis, many hospitals operated close to ICU capacities [566], raising the - ethically difficult - question on how to justly allocate capacities in acute resource scarcity [567].
The CDSS developed in **Chapter 2 and 3** focuses on the prediction of clinical outcomes essential in patient management, namely the estimated 30-day mortality of NSTI patients and the likely disease severity for COVID-19. Consequently, it can be utilised to predict supply shortages and facilitate timely patient transfers if necessary. This proactive approach helps mitigate the need for immediate triage, potentially averting or delaying such measures altogether.
Considering the large local difference in treatment guidelines and resources among care facilities, the localisation of pre-trained models presented in **Chapter 4** could prove invaluable. It enables fine-tuning of models to provide suggestions aligned with the practical

realities of each hospital. Such customisation significantly enhances the clinical utility of CDSS in patient management, ensuring recommendations are tailored to the resources available within each facility.

9.1.5 Novel Treatments and Diagnostic Biomarker Sets

Next to effectively using the resources and options available to us today, the healthcare economy has a big interest in the constant development of novel diagnostic tools and treatment options. To avert another crisis comparable to the antibiotics resistance crisis, which can be largely attributed to the lack of new drug development [568], there is a continuous endeavour to uncover disease mechanisms that can be exploited to our advantage. In recent decades, however, we observed an increase in costs of drug development [569] and a decline in the efficiency of pharmaceutical research and development, which can be attributed to increasing complexity of drug targets [570, 571].

Today, the capability of AI to relate clinical outcomes to complex, multi-dimensional data patterns, makes it an invaluable tool in the life sciences. I argue that failing to understand the rationale behind AI model decision-making would be a missed opportunity, hindering the advancement of novel drugs and biomarkers. However, the interpretation of AI models is neither trivial nor straightforward, especially for researchers new to computational biology. Every data type and model architecture offers multiple interpretation strategies, each based on different sets of assumptions. Despite the plethora of reviews delineating the various possibilities of interpreting models [333, 352, 572, 573], first-time deep learning users may easily feel overwhelmed with the combinatorial possibilities presented to them. With the aim to guide fellow researchers and unravel the convoluted field that is explainable genomics, **Chapter 6** is our attempt to capture the current state of the field by quantifying which interpretability strategies are commonly paired with which data types, model architectures, and scientific questions. Our work distinguishes itself by providing a meticulously curated and fully interactive overview table, allowing readers to retrieve studies matching their own research questions, data, or preferred model architecture and provide concrete inspiration on which interpretability methods to test.

In **Chapter 7** I aimed to showcase the possibilities explainable AI provides for hypothesis generation, focusing on its application in the field of epigenetics, in particular DNA methylation. Obstructed by a large number of methylated sites, fully understanding the organisation and interactions within our methylomes remains challenging [464]. Our research proved that representation learning can be successfully used to filter out redundant and correlated signals, significantly reducing the volume of information to analyse. A post-hoc explainability strategy subsequently revealed groups of methylated sites with

biological significance. However, these findings couldn't be readily explained by common correlation patterns, spatial proximity, or biological pathways. This suggests that the proposed framework could serve as a valuable tool for generating novel hypotheses regarding CpG interactions, which can then be investigated further experimentally.

Precision medicine is an inherently interdisciplinary discipline, necessitating the integration of expertise from various fields, including genomics, informatics, medicine, and ethics to develop novel diagnostic tools and treatment options. Therefore, training the next generation of precision medicine experts thus requires the development of cross-disciplinary, international, and education-focused doctoral programs [574]. The European Commission has recognised this need and funded TranSYS, an Innovative Training Network (ITN) for precision medicine. As a participant in TranSYS, my fellow early-stage researchers and I critically evaluated the training and collaboration experiences within our ITN, which are presented in **Chapter 8** of this thesis. Our quantitative analysis revealed the significant scientific output enabled by TranSYS, ranging from biomarker discovery and molecular mechanisms to patient stratification and the deployment of AI-based computational tools. Also, we found substantial personal improvements in scientific, professional, and social skills among young researchers, clearly linked to interactions within the network. However, we also identified several hurdles and shortcomings, such as a lack of communication among supervisors, insufficient managerial resources, and inadequate training in soft skills related to personal well-being.

Taken together, we believe the insights generated through this work can add to the continuous development of the Marie Skłodowska-Curie Action (MSCA) Doctoral Training Networks and improve the training of precision medicine researchers following in our footsteps.

9.2 Barriers to Implementing Precision Medicine

Given the abundance of perspective and opinion articles extolling the impact of precision medicine, critical minds may ask themselves why every hospital worldwide has not yet adopted precision medicine practices. After all, other AI models such as OpenAI's ChatGPT have managed to take the Western world by storm, acquiring 1 million users just 5 days after their launch [575, 576].

So why are we still slacking to incorporate CDSS into medical routines or do not take into account genomic information in clinical decisions? Throughout the research leading up to this thesis, I have asked myself these exact questions and realised that the practical implementation of PM is hindered by a number of fundamental shortcomings with respect

to data infrastructure and abundance, as well as model development, explainability, and transferability (Figure 9.2).

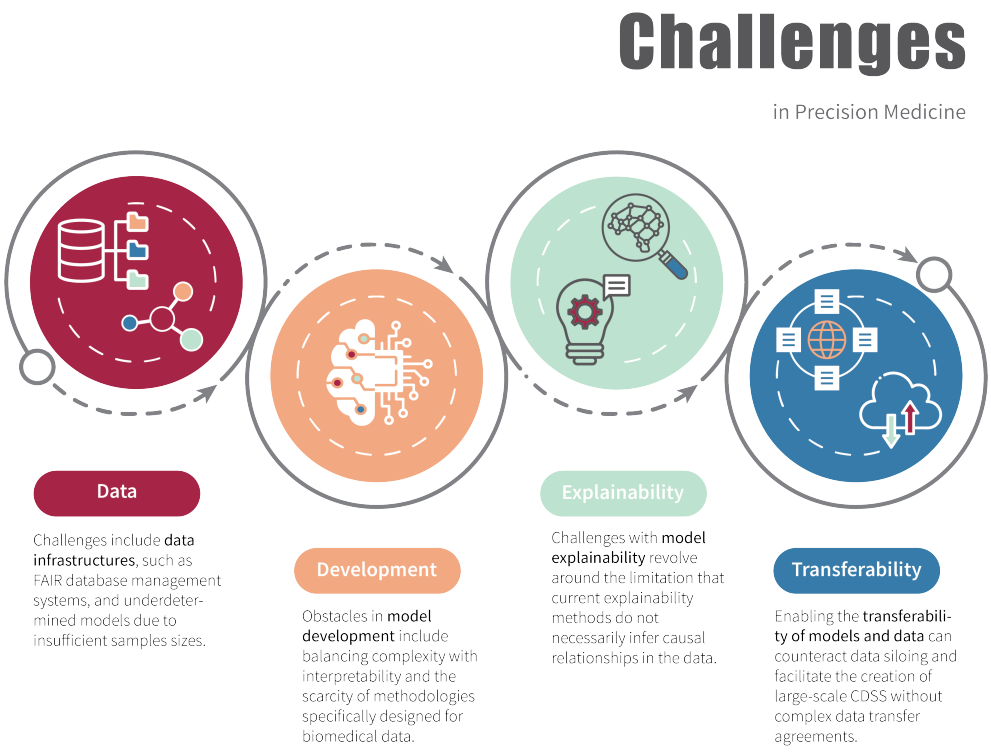


Figure 9.2: **Challenges in precision medicine identified in this Thesis.** Presented are shortcomings with respect to Data, Model Development, Model Explainabiliy, and Data and Model Transferability that were encountered through the work associated with this Thesis. Icons retrieved from The Noun Project (CC BY 3.0)

9.2.1 Data Challenges: Is Big Data Too Big, Not Big Enough, or Both Simultaneously?

Currently, a large amount of research efforts are invested in the development of high-performance inference engines of CDSS. While inference engines - the AI-powered models that process patient information and generate decisions - are undoubtedly the core of every CDSS, they can not make predictions without a comprehensive knowledge base to extract data from.

The development of a robust and versatile knowledge base is a crucial aspect that presents significant challenges. For example, how to organise and store heterogeneous clinical data,

such as images, laboratory results, and handwritten notes, without hindering users from extracting or entering data efficiently? The significance of a well-designed knowledge base in a CDSS, or a hospital's overall system, is exemplified by the recent problematic deployment of Epic's electronic health record (EHR) system in Europe.

The US-based Epic Systems Corporation, commonly known as Epic, develops and services relational database management systems for whole hospitals. Their recent launch in multiple European countries, including the UK, Finnland, Denmark, and Norway was characterised by staff reporting chaos due to a lack of training, unstable software, and a UI so convoluted, it endangered patient safety. Launches culminated in large protests by hospital staff, petitions to remove the system, and reports of doctors experiencing stress-related health issues due to the new IT system, that made them consider to quit their job [577].

Precision medicine's goal of creating a more transparent healthcare system requires us to rethink how I store data in order for it to be machine-readable, interoperable between systems, and reusable. While the system developed by Epic promotes data siloing, going as far as charging the sharing of data with users of competitors' software [578], more forward-thinking infrastructures endorse the FAIR (Findability, Accessibility, Interoperability, and Reusability) Data Principles [579, 580]. Combining FAIR principles with semantic web technologies allows us to construct databases that effectively link diverse data modalities in a structured manner [581–583]. This fosters a circular data economy, ensuring the long-term sustainability of CDSS and other precision medicine tools. As increasingly more patient-specific data is generated, which could be leveraged for the patient's benefit, there is a pressing need to shift researchers' focus from developing inference machines to establishing FAIR semantic databases in clinical settings.

The work encompassed in this thesis entailed the development of methodologies for a variety of data types, diseases, and research questions. Yet, irrespective of the specific focus, we encountered a similar challenge throughout the field: underdetermined systems. An underdetermined system refers to a system, in which the number of samples is insufficient relative to the number of parameters being estimated. This means that there may be multiple possible solutions that fit the available data equally well, or the model may be unable to accurately capture the underlying relationships in the data due to the lack of information [584]. With the millions of single nucleotide polymorphisms genomics data yields, or the large number of trainable parameters in a deep neural network, working with biomedical data implies that systems are more commonly than not underdetermined. While often overlooked, this harbours a range of risks, including a lack of unique solutions, the difficulty of models to generalise, or a large uncertainty to predictions [65].

With large language models (LLM) getting increasingly adopted into computational biology, this problem is prone to increase. LLM may require millions of tokes to sufficiently train [585], which is not easily obtained from the biomedical dataset. The recent surge in the application of LLM to various biomedical issues underscores another inherent challenge in the field: the scarcity of original methodologies. Given that LLM stems from natural language processing (NLP), they are designed to handle text-like input formats, which are rarely encountered in biological data.

Currently, the approach of using LLM with biomedical data involves treating it as text [586]; while this may not be far from the truth for genomic and protein sequences, other (numerical) data types, such as gene and protein expression or DNA methylation may not be as interpreted as naively.

Despite the trend of Big and even Bigger Data continuing, the amount of information available in biomedical domains will never be equal to the wealth of input available for other NLP or imaging tasks, due to the resource-intensive process of collection, annotating, and handling sensitive data. Therefore, instead of developing many individual, underdetermined systems, more communal effort should be put into data and model sharing using transfer learning (more on that in section 9.2.4).

9.2.2 Challenges in Model Development: The Ever-Increasing Complexity of Models

With the latest advancements in LLM and quantum computing expected to provide close to unlimited computing power in the near future, the field of computational biology seems to be in a constant race to create more extensive and sophisticated frameworks.

The prevailing belief is that we require increasingly complex frameworks to effectively model biomedical data. This is supported by the often-cited trade-off between model interpretability and complexity, which implies that inherently interpretable models, such as logistic regression, suffer in performance with respect to larger models with more parameters such as deep learning models. However, there is evidence to believe this to be nothing but a wrong dichotomy [587, 588].

Recent studies have proven that some computational problems might not benefit from the large modelling capacities complex models like LLM possess; in simple tasks, such as cell type annotation in single-cell data LLM do not outperform logistic regression [589]. Some studies have shown that logistic regression models, when enhanced through feature extraction, exhibit predictive performances comparable to those of more complex supervised models like gradient-boosted trees [590]. Moreover, it has been reported that the significant impact of hyperparameter tuning on performance may even surpass the

variability observed between different models during model selection [591].
To effectively transition precision medicine from research to clinical practice, researchers should redirect their focus away from rapidly adopting new methods and generating immediate results. Instead, we should prioritise the exploration of the plethora of methodologies available to us, and aim to fully leverage their potential.
As a concrete example, when comparing different machine learning models, commonly their performance is compared, and the best performing one is selected. "Performance" in this respect is commonly limited to metrics capturing model accuracy, not taking into account other important factors, such as model interpretability or complexity. In concordance with others [90, 592], I argue that this is a significant short-slightness in the field, as nowadays models intended to be of clinical use are required to be more than just accurate. The advantages a complex deep learning model brings over an inherently interpretable machine learning model must be carefully weighed, and when in doubt, the more simplistic model should be preferred. In addition thereof, computational biology has to stop blindly adopting methodologies from other fields and start developing solutions tailored to and utilising the uniqueness of biomedical data. Recently, for instance, an expert on causal inference suggested that the biomedical sciences possess a range of interventional and observational tools invaluable for causality research, a feature seldom found in other disciplines [593].

9.2.3 Challenges with Respect to Model Explainability: Correlation, Causation, or Randomness?

With experimental validations remaining expensive and often challenging, it is more often than not unclear if the statistical dependencies models capture have "true" biological meaning, or are merely co-correlations of causal biology, technical artefacts, such as measurement biases, confounders, or even random associations. A multitude of biomarker combinations can be constructed to separate the same patient groups [594] and this raises further questions concerning the causality underlying the biological associations uncovered by explainable AI. While it can be seen as a major advantage of AI models to not require a detailed understanding of cause-effect relationships or biological mechanisms to start modelling data, introducing causality or mechanistic modelling into black box models may represent the next step in the interpretation of complex models (Figure 9.3).
Causality research aims to answer questions central to precision medicine, such as "What is the best treatment to recommend for this patient?", by modelling how variables influence each other in a cause-effect relationship. The treatment effect, for instance, is the difference between two outcomes: the factual outcome (e.g. treatment A received) and the

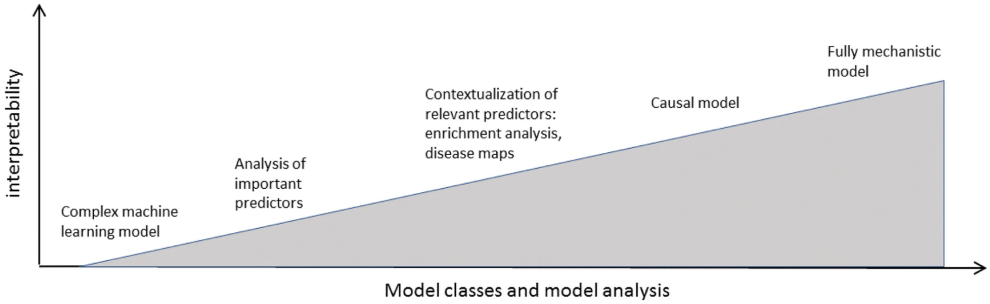


Figure 9.3: **Different classes of machine learning models and their interpretability.** Reprinted from 'From hype to reality: data science enabling personalized medicine', by Holger Fröhlich *et al.*, in *BMC medicine*, 2018, Springer [595].

counterfactual one (treatment B received); logically, only one of them can be observed in reality. While traditional machine learning focusing solely on association can infer which treatment similar patients have received, therefore giving suggestions, they can not estimate the effect different treatments have on patients. Causal models attempt to fill this gap and reason about hypothetical scenarios in which different actions could have been taken [596]. Commonly, causal models are presented as directed acyclic graphs where nodes present biomedical information, such as symptoms, diseases, and risk factors, and edges denote their causal relationship [593, 597].

The main challenge of working with causal models is their construction, as the underlying causal graph can only be recovered up to a certain degree from real-world observational data [598]. Missing information must be filled by making assumptions, which need to constantly be re-evaluated and require in-depth domain knowledge to be constructed. The possibilities causal modelling promises, however, are significant: it can be used to identify variables with influence on patient outcomes [599], identify novel drug targets, estimate toxicity and dosing of novel drugs, select the most fitting patients for clinical trials [593], or act as a *in-silico* patient representation to determine the most promising interventions [597].

While causal modelling tries to infer relationships without the need to understand the whole underlying system, **mechanistic modelling** takes these ambitions further and aims at representing biological systems based on known physiological mechanisms using exact mathematical equations. With a more bottom-up *in-silico* approach based on the understanding of the mechanism of action [600], examples of mechanistic models include pharmacokinetic [601, 602] and metabolic modelling [603, 604]. However, the construction of mechanistic models is challenging, as it requires knowledge of disease-driving mechanisms across scales [595].

However, the widespread use of fully mechanistic models in the near future is rather unlikely given the number of only partially understood biological mechanisms, available knowledge can be implemented into machine learning frameworks forming hybrid or grey box models [112]. Although hybrid models have shown success in the past [112], more case studies are needed to generate interest in this area of research.

9.2.4 Transferability Challenges: The Future Is Decentralised

As demonstrated throughout this thesis, the quality of a CDSS mainly hinges on the amount, quality, accessibility, and representativeness of data. Overall, the lack of resources dedicated to data interoperability and sharing hinders the exploration of CDSS usability and consequently undermines the establishment of trust in this upcoming technology [546–549]. As the collection of such amounts of diverse data for sure exceeds the capabilities and finances of individual institutions, data sharing holds the key to realise large-scale, and therefore powerful, CDSS.

This is of special importance for rare diseases characterised by their extremely low number of incidences, such as NSTI [605]. However, these centralised learning strategies, for which all data is transferred to a central server which does calculations, are embedded in privacy, ownership, and regulatory issues that prevent combining medical data into traditional centralised storage. Currently, data transfers from the European Union (EU) to the United States (US) for medical research are impeded, as the European (GDPR) imposes strict rules on the transfer of personal data outside the EU to ensure adequate levels of protection [605]. The US does not automatically meet this high standard, primarily due to differences in data protection laws and practices, resulting in a halting of transfers. However, even for data transfers within the EU, the arrangement of data agreements can take weeks to months [606].

Combating these problems, the recently proposed counterpart of centralised learning, termed decentralised learning, aims to provide a privacy-preserving alternative which does not require sensitive information to be shared. A decentralised learning strategy, such as federated learning avoids the sharing of medical data by utilising a central server only to store the model and its parameters. Cyclically, the servers share respective parameters with each client, for example, each hospital, which trains a local model on their locally stored data. After training, only the updated model parameters are returned to the central server, which aggregates the model parameters from various clients and sends the aggregated parameters back to the clients [607].

Even though decentralised learning strategies need to overcome several practical obstacles, from technical problems such as communication failures or large energy consumption

[608] to problems specific to biomedical data such as differences in data annotation, noise, and heterogeneity [609], it is clear that this nascent field is the future of clinical research. Its potential to maintain data privacy while increasing the generalisability and power of models, holds the key to enabling large-scale international collaboration.

9.3 The Digitalisation of Hospitals as an Opportunity

Healthcare infrastructures currently invest millions into fast-paced digitalisation programs; Berlin's Charité university hospital aims to be fully digitalised by 2030 [610], while the Germany-wide digitalisation of hospitals is estimated to take a yearly 2 billion euros of investment [611]. I believe this ambitious transitioning phase is the opportunity to address many of the shortcomings outlined in the previous section, especially those related to data infrastructures.

For instance, the investment into computational resources will not only enable in-house research, advanced analytics, and federated learning infrastructures but pave the way to the establishment of a circular data design using FAIR database management systems. "Computational infrastructures" however do not only comprise the acquisition of machines, but include the employment of highly-educated and specialised staff like data semantics and ontology experts, computer scientists, and biomedical data engineers. Additionally, with next-generation sequencing becoming cheaper by the year, hospitals now are in a position to acquire and implement omics data into clinical workflows. Throughout this thesis, I continuously focused on providing explanatory models, with the idea for medical staff to foster trust in AI models and increase societal acceptance.

Aversions to these novel and complex innovations, however, may be unrelated to a lack of insight but rather stem from a general lack of digital competencies of staff [612]. Fortunately, as digital-native generations like Generation Z enter the workforce and Millennials advance into higher positions, while the baby boomer generation gradually retires, educating end-users on the responsible use of AI becomes more feasible. Naturally, combining a native digital understanding with education on the opportunities, limitations, and pitfalls of AI will significantly enhance trust in these innovations, potentially leading to the disappearance of aversions without extensive interventions.

As a summary, I believe that the ongoing digitalisation of hospitals presents a unique opportunity for precision medicine advocates to shape the vision of healthcare 4.0. This offers a realistic prospect of bridging the gap between computational biology research and clinical care, thereby moving closer to the vision of a more integrated and personalised healthcare system.

In conclusion, this thesis has aimed at exploring the integration of precision medicine concepts into clinical practice, leveraging computational tools as the means to achieve this objective.

The CDSS developed for Necrotising Soft Tissue Infections and COVID-19 not only enhance clinical decision-making but also provide transparent explanations for the decisions made, thereby strengthening the reliability and trustworthiness of the system.

Moreover, the developed deep learning framework, capable of simultaneously integrating multidimensional genomic data and correcting for potential confounders, presents a fair patient stratification tool. This tool holds promise in enhancing treatment effectiveness and minimising side effects by identifying the patient groups most likely to benefit from specific treatments.

Across various projects, emphasis has been placed on the explainability of deep learning methods. The research presented in this thesis included a quantitative analysis on commonly used methodologies to interpret deep learning genomics frameworks. These methods have proven advantageous in navigating high-dimensional data, facilitating hypothesis generation beyond the capabilities of traditional analysis methods.

With the goal of contributing to the scientific community, this thesis includes a critical assessment of the training and educational proficiency of Innovative Training Networks (ITNs). By analysing the personal and professional experiences of young researchers within ITNs and highlighting the significant drawbacks of these networks, this thesis aims to pave the way for improvements in ITNs and empower the next generation of young researchers to realise their full potential.

Throughout the work underlying this thesis several barriers hindering the practical implementation of precision medicine have been encountered and identified. These barriers encompass challenges related to data infrastructure, model development and explainability, as well as the transferability of models and data. I argue that with the ongoing digitalisation of hospitals, the time has come to address the highlighted problems, particularly concerning data management and sharing. By addressing these issues proactively, we can envisage a future where Clinical Decision Support Systems and precision medicine concepts transition from research to clinical practice.

9.4 Supplementary Notes

Acknowledgements

Language, expression, and grammar in this chapter were polished using ChatGPT to improve readability.

Summary

In an era marked by an unprecedented wealth of information, technological possibilities, and computational resources, Artificial Intelligence (AI) stands at the epicentre of the digital revolution reshaping Western societies. In addition, scientific advancements of the last decades have broadened our understanding of health, the human body, and diseases, leading to the growing consensus that I need to modernise our healthcare systems and harness the technological advancements readily available to us. These societal needs are reflected in the concept of precision medicine, with its overarching goal to transform healthcare systems towards transparency, data-driven practices, and personalised approaches.
This thesis seeks to advance the principles of precision medicine and facilitate this transformation by developing artificial intelligence models that utilise diverse biomedical data, emphasising explainability and fairness. The concrete goals of this thesis therefore are as follows: (I) develop machine learning-based clinical decision support tools for Necrotising Soft Tissue Infections (NSTI) and COVID-19, (II) design fair stratification tools for cancer patients that integrate information from diverse biological layers, and (III) utilise explainable deep learning algorithms to explore the complexities of the human epigenome.

The first goal, the development of clinical decision support tools, is addressed in **Part 1**, which explores how we can utilise precision medicine concepts to enhance the management and care of patients suffering from NSTI (**Chapter 2 and 4**) and COVID-19 (**Chapter 3**).
Chapter 2 explores the possibility of utilising machine-learning methods, namely Random Forest Classifiers, to estimate the risk of mortality of NSTI patients. However, the chapter also provides an overview of relevant clinical needs in NSTI care, derived from qualitative interviews with clinicians from different departments. These interviews yielded 24 unique clinical questions, which served as the base for further projects, such as the work presented in **Chapter 4**. The risk of 30-day mortality could be accurately estimated using clinical parameters available in the data set on the first 24 h following ICU admission (ROC AUC:

0.91 (95% CI, 0.88-0.96)), thereby surpassing the performance of clinical scoring systems commonly utilised for the same purpose. Analysis of feature importances revealed that the selected variables are highly clinically relevant, as they are linked to septic shock, a frequent cause of NSTI mortality, further increasing trust in the system. This chapter highlights the potential of machine learning models to enhance clinical decision-making, resource allocation, and early communication with patients and families.

In **Chapter 3**, the knowledge gained from developing a CDSS for NSTI was applied to COVID-19. This chapter aims to evaluate the effects of COVID-19 on intracellular cholesterol biosynthesis and develop a predictive model for COVID-19 severity. In an attempt to assess the potential new biomarkers for disease monitoring, the benefits of measuring metabolic pathways, specifically endogenous cholesterol biosynthesis, were explored. Several machine learning models were trained on a combination of clinical data and sterol intermediates to predict disease severity. These models ultimately outperformed the COVID-GRAM, a comparable clinical model developed for the same purpose. This chapter demonstrates the capability of AI-powered CDSS to seamlessly integrate data from diverse domains into a unified model and directly evaluate the predictive power of new biomarkers.

Chapter 4 revisits the clinical needs associated with NSTI by exploring various clinical endpoints, such as the need for amputation, anticipated wound size, estimated length of ICU stay, and the presence of specific pathogens, such as Group A Streptococcus (GAS). By attempting to predict each of these endpoints through machine learning models, it was found that GAS aetiology can be predicted before the first surgery, significantly earlier than current methods allow. However, predicting other endpoints besides GAS efficiently was not possible due to biased outcomes influenced by individual surgeons or local guidelines, and the need for more longitudinal data, highlighting shortcomings in current NSTI research. Overall, the findings presented in this chapter could enable targeted interventions for GAS earlier in the NSTI disease course, demonstrating the high clinical relevance of results generated by a CDSS.

Part 2 of this thesis addresses the second goal formulated, namely the design of fair stratification tools for cancer patients that integrate information from diverse biological layers. While grouping patients using high-dimensional, multi-omics data is common, patient stratification is practically often limited by the impact of confounders like batch effects, age, or sex on clustering.

Chapter 5 therefore introduces novel variational autoencoder-based frameworks on multi-omics data capable of mitigating these confounders while preserving biological patterns.

Four different deconfounding strategies are presented and compared using data with simulated confounding effects. This chapter illustrates the power of deep learning not only to integrate diverse omics measurements into a unified framework but also to effectively handle confounding effects and recover biologically driven clustering structures. It accurately identifies patient labels and uncovers meaningful pathological associations, offering guidance for meaningful patient stratification in precision medicine.

The focus of **Part 3** lies in demonstrating how explainable deep learning algorithms can be used to explore the complexities of the human epigenome and drive hypothesis generation in genomics.
Chapter 6 addresses the fact, that while the interpretation of deep learning models is challenging, it is essential for understanding biological mechanisms and establishing trust in healthcare applications. Therefore, this chapter offers an overview of interpretable deep learning solutions in functional genomics. It includes a quantitative exploration of framework design choices concerning genomics data characteristics, neural network architectures, and interpretation strategies. Through the evaluation of the genomics field using predefined criteria, common solutions are identified, and opportunities for developing more interpretable models are uncovered. The chapter aims to serve as guidance for genomic researchers looking to integrate interpretable deep learning models into their methodologies.
Chapter 7, on the other hand, provides a concrete showcase of how explainable AI can be utilised to drive hypothesis generation in epigenomics. With traditional methods often falling short in understanding the relationships among DNA methylation sites due to the complexity and high dimensionality of the data, deep learning algorithms like autoencoders offer promising solutions. This chapter introduces mEthAE, a chromosome-wise autoencoder designed to interpret and reduce methylation data dimensions effectively. In an attempt to go beyond mere data compression, an explainability approach is presented, which successfully identifies clusters of biologically relevant DNA methylation sites exhibiting strong connections across the reduced latent space. Interestingly, these connected sites do not show correlation patterns or physical proximity on chromosomes, suggesting the detection of complex, long-range, non-linear interactions by our autoencoder, suggesting an interaction pattern largely uncharacterised in current epigenetic research. This chapter demonstrates the potential of explainable deep learning frameworks in unravelling the complexities of (epigenetic) data and proposes new hypotheses in genomics.

The final part of this thesis, **Part 4**, explores the interdisciplinary nature of precision

medicine. Addressing the challenges associated with the paradigm shift in disease prevention and healthcare requires integrating expertise from genomics, informatics, medicine, and ethics. Training the next generation of precision medicine experts, therefore, necessitates developing cross-disciplinary, international doctoral programs focused on education. **Chapter 8** follows this vision and introduces the training and collaboration experiences of early-stage researchers within TranSYS, an Innovative Training Network (ITN) for precision medicine funded under the EU's Horizon Europe program. Training and collaboration aspects within the ITN are examined, revealing significant improvements in young researchers' scientific, professional, and social skills. Case studies highlight the benefits of collaborative environments and innovative scientific exchange within TranSYS. Despite challenges like balancing stakeholder interests, the ITN was found to foster positive growth within young researchers. Overall, this chapter provides insight into the scientific topics and training of the next generation of precision medicine experts.

The thesis ends with a **Discussion** that summarises the key findings and situates them within the broader context of precision medicine. It shows the value of the work presented in this thesis concerning different goals of precision medicine, including the need for accurate decision making, explainability, effective treatment outcomes, optimally allocating resources, and finding novel biomarkers and treatment. A significant portion of the Discussion is devoted to highlighting the open challenges that need to be addressed to integrate precision medicine into routine clinical practice. Lastly, the chapter presents a perspective on the opportunities crucial for shaping the future of the field.

Bibliography

1. Fierz, W. Challenge of Personalized Health Care: To What Extent Is Medicine Already Individualized and What Are the Future Trends? eng. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* **10,** RA111–123 (2004).
2. Hope, B. The Greatest Benefit to Mankind: A Medical History of Humanity From Antiquity to the Present. *BMJ : British Medical Journal* **316,** 713 (1998).
3. *A Dictionary of Public Health* en_US (ed Last, J. M.) ISBN: 978-0-19-516090-1 (Oxford University Press, 2007).
4. Hempel, S. *The Medical Detective: John Snow, Cholera and the Mystery of the Broad Street Pump* (Granta Books, 2014).
5. Eddy, D. M. *Clinical Decision Making: From Theory to Practice: A Collection of Essays from the Journal of the American Medical Association* (American Medical Association, 1996).
6. Bovis, F., Carmisciano, L., Signori, A., Pardini, M., Steinerman, J. R., Li, T., Tansy, A. P. & Sormani, M. P. Defining Responders to Therapies by a Statistical Modeling Approach Applied to Randomized Clinical Trial Data. *BMC Medicine* **17,** 113. doi:10.1186/s12916-019-1345-2 (2019).
7. McLeod, H. L. & Evans, W. E. Pharmacogenomics: Unlocking the Human Genome for Better Drug Therapy. *Annual review of pharmacology and toxicology* **41,** 101–121 (2001).
8. Collins, F. S. & McKusick, V. A. Implications of the Human Genome Project for Medical Science. *JAMA* **285,** 540–544. doi:10.1001/jama.285.5.540 (2001).

9. Kravitz, R. L., Duan, N. & Braslow, J. Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages. en. *The Milbank Quarterly* **82,** 661–687. doi:10.1111/j.0887-378X.2004.00327.x (2004).

10. Kosorok, M. R. & Laber, E. B. Precision Medicine. *Annual review of statistics and its application* **6,** 263–286 (2019).

11. Genosalut_Palma. *Personalised Medicine: What Is It and What Are Its Objectives* en-GB. 2022.

12. Phillips, C. J. Precision Medicine and Its Imprecise History. en. *Harvard Data Science Review* **2.** doi:10.1162/99608f92.3e85b56a (2020).

13. Schleidgen, S., Klingler, C., Bertram, T., Rogowski, W. H. & Marckmann, G. What Is Personalized Medicine: Sharpening a Vague Term Based on a Systematic Literature Review. *BMC medical ethics* **14,** 1–12 (2013).

14. Nimmesgern, E., Benediktsson, I. & Norstedt, I. Personalized Medicine in Europe. *Clinical and Translational Science* **10,** 61–63. doi:10.1111/cts.12446 (2017).

15. Fröhlich, H. *et al.* From Hype to Reality: Data Science Enabling Personalized Medicine. *BMC Medicine* **16,** 150. doi:10.1186/s12916-018-1122-7 (2018).

16. Vogenberg, F. R., Isaacson Barash, C. & Pursel, M. Personalized Medicine: Part 1: Evolution and Development into Theranostics. eng. *P & T: A Peer-Reviewed Journal for Formulary Management* **35,** 560–576 (2010).

17. Mathur, S. & Sutton, J. Personalized Medicine Could Transform Healthcare. *Biomedical reports* **7,** 3–5 (2017).

18. Norrby-Teglund, A., Svensson, M. & Skrede, S. *Necrotizing Soft Tissue Infections: Clinical and Pathogenic Aspects* en. ISBN: 978-3-030-57616-5 (Springer Nature, 2020).

19. Madsen, M. B. *et al.* Patient's Characteristics and Outcomes in Necrotising Soft-Tissue Infections: Results from a Scandinavian, Multicentre, Prospective Cohort Study. eng. *Intensive Care Medicine* **45,** 1241–1251. doi:10.1007/s00134-019-05730-x (2019).

20. Bisno, A. L., Cockerill III, F. R. & Bermudez, C. T. The Initial Outpatient-Physician Encounter in Group A Streptococcal Necrotizing Fasciitis. *Clinical infectious diseases* **31,** 607–608 (2000).

21. van Stigt, S. F. L., de Vries, J., Bijker, J. B., Mollen, R. M. H. G., Hekma, E. J., Lemson, S. M. & Tan, E. C. T. H. Review of 58 Patients with Necrotizing Fasciitis in the Netherlands. eng. *World journal of emergency surgery: WJES* **11,** 21. doi:10.1186/s13017-016-0080-7 (2016).

22. Audureau, E., Hua, C., de Prost, N., Hemery, F., Decousser, J. W., Bosc, R., Lepeule, R., Chosidow, O., Sbidian, E. & Henri Mondor Hospital Necrotizing Fasciitis group. Mortality of Necrotizing Fasciitis: Relative Influence of Individual and Hospital-Level Factors, a Nationwide Multilevel Study, France, 2007-12. eng. *The British Journal of Dermatology* **177,** 1575–1582. doi:10.1111/bjd.15615 (2017).

23. Tom, L. K., Maine, R. G., Wang, C. S., Parent, B. A., Bulger, E. M. & Keys, K. A. Comparison of Traditional and Skin-Sparing Approaches for Surgical Treatment of Necrotizing Soft-Tissue Infections. eng. *Surgical Infections* **21,** 363–369. doi:10.1089/sur.2019.263 (2020).

24. Al-Qurayshi, Z., Nichols, R. L., Killackey, M. T. & Kandil, E. Mortality Risk in Necrotizing Fasciitis: National Prevalence, Trend, and Burden. *Surgical Infections* **21,** 840–852 (2020).

25. Horn, D. L., Shen, J., Roberts, E., Wang, T. N., Li, K. S., O'Keefe, G. E., Cuschieri, J., Bulger, E. M. & Robinson, B. R. H. Predictors of Mortality, Limb Loss, and Discharge Disposition at Admission among Patients with Necrotizing Skin and Soft Tissue Infections. eng. *The Journal of Trauma and Acute Care Surgery* **89,** 186–191. doi:10.1097/TA.0000000000002636 (2020).

26. Fagerdahl, A.-M., Knudsen, V. E., Egerod, I. & Andersson, A. E. Patient Experience of Necrotising Soft-Tissue Infection from Diagnosis to Six Months after Intensive Care Unit Stay: A Qualitative Content Analysis. eng. *Australian Critical Care: Official Journal of the Confederation of Australian Critical Care Nurses* **33,** 187–192. doi:10.1016/j.aucc.2019.02.001 (2020).

27. Urbina, T. *et al.* Long-Term Quality of Life in Necrotizing Soft-Tissue Infection Survivors: A Monocentric Prospective Cohort Study. *Annals of Intensive Care* **11,** 102. doi:10.1186/s13613-021-00891-9 (2021).

28. Suijker, J., de Vries, A., de Jong, V. M., Schepers, T., Ponsen, K. J. & Halm, J. A. Health-Related Quality of Life Is Decreased After Necrotizing Soft-Tissue Infections. *Journal of Surgical Research* **245,** 516–522. doi:10.1016/j.jss.2019.07.097 (2020).

29. Suijker, J., Stoop, M., Meij-de Vries, A., Pijpe, A., Boekelaar, A., Egberts, M. & Van Loey, N. The Impact of Necrotizing Soft Tissue Infections on the Lives of Survivors: A Qualitative Study. en. *Quality of Life Research* **32,** 2013–2024. doi:10.1007/s11136-023-03371-8 (2023).

30. Erichsen Andersson, A., Egerod, I., Knudsen, V. E. & Fagerdahl, A.-M. Signs, Symptoms and Diagnosis of Necrotizing Fasciitis Experienced by Survivors and Family: A Qualitative Nordic Multi-Center Study. *BMC infectious diseases* **18,** 1–9 (2018).

31. Hedetoft, M., Madsen, M. B., Madsen, L. B. & Hyldegaard, O. Incidence, Comorbidity and Mortality in Patients with Necrotising Soft-Tissue Infections, 2005–2018: A Danish Nationwide Register-Based Cohort Study. *BMJ open* **10,** e041302 (2020).

32. Stevens, D. L. & Bryant, A. E. Necrotizing Soft-Tissue Infections. eng. *The New England Journal of Medicine* **377,** 2253–2265. doi:10.1056/NEJMra1600673 (2017).

33. Peetermans, M., de Prost, N., Eckmann, C., Norrby-Teglund, A., Skrede, S. & De Waele, J. J. Necrotizing Skin and Soft-Tissue Infections in the Intensive Care Unit. eng. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **26,** 8–17. doi:10.1016/j.cmi.2019.06.031 (2020).

34. Medina, L. M. P. *et al.* Discriminatory Plasma Biomarkers Predict Specific Clinical Phenotypes of Necrotizing Soft-Tissue Infections. en. *The Journal of Clinical Investigation* **131.** doi:10.1172/JCI149523 (2021).

35. Jahagirdar, S. *et al.* Analysis of Host-Pathogen Gene Association Networks Reveals Patient-Specific Response to Streptococcal and Polymicrobial Necrotising Soft Tissue Infections. *BMC Medicine* **20,** 173. doi:10.1186/s12916-022-02355-8 (2022).

36. Page, J., Hinshaw, D. & McKay, B. In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market–The Man with the First Confirmed Infection of the New Coronavirus Told the WHO Team That His Parents Had Shopped There. *The Wall Street Journal* **26** (2021).

37. WHO. *WHO* 2024. https://covid19.who.int/.

38. Wikipedia. *List of Epidemics and Pandemics — Wikipedia, the Free Encyclopedia* 2024. http://en.wikipedia.org/w/index.php?title=List%5C%20of%5C%20epidemics%5C%20and%5C%20pandemics%5C&oldid=1208355218.

39. Teodori, L., Osimani, B., Isidoro, C. & Ramakrishna, S. Mass versus Personalized Medicine against COVID-19 in the "System Sciences" Era. en. *Cytometry Part A* **101,** 995–999. doi:10.1002/cyto.a.24662 (2022).

40. Ledford, H. How Common Is Long COVID? Why Studies Give Different Answers. en. *Nature* **606,** 852–853. doi:10.1038/d41586-022-01702-2 (2022).

41. Arish, M. & Naz, F. Personalized Therapy: Can It Tame the COVID-19 Monster? *Personalized Medicine* **18,** 583–593. doi:10.2217/pme-2021-0077 (2021).

42. Teodori, L., Osimani, B., Isidoro, C. & Ramakrishna, S. Mass versus Personalized Medicine against COVID-19 in the "System Sciences" Era. en. *Cytometry Part A* **101,** 995–999. doi:10.1002/cyto.a.24662 (2022).

43. Soni, D., Van Haren, S. D., Idoko, O. T., Evans, J. T., Diray-Arce, J., Dowling, D. J. & Levy, O. Towards Precision Vaccines: Lessons From the Second International Precision Vaccines Conference. *Frontiers in Immunology* **11** (2020).

44. Bakkerus, L. & Pickkers, P. Personalized Medicine in COVID-19. *Intensive Care Medicine* **48,** 1607–1610. doi:10.1007/s00134-022-06908-6 (2022).

45. Radanliev, P., De Roure, D., Walton, R., Van Kleek, M., Montalvo, R. M., Santos, O., Maddox, L. & Cannady, S. COVID-19 What Have We Learned? The Rise of Social Machines and Connected Devices in Pandemic Management Following the Concepts of Predictive, Preventive and Personalized Medicine. *The EPMA Journal* **11,** 311–332. doi:10.1007/s13167-020-00218-x (2020).

46. Filip, R., Gheorghita Puscaselu, R., Anchidin-Norocel, L., Dimian, M. & Savage, W. K. Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: A Review of Pandemic Measures and Problems. *Journal of Personalized Medicine* **12,** 1295. doi:10.3390/jpm12081295 (2022).

47. Scerri, M. & Grech, V. Artificial Intelligence in Medicine. *Early human development* **145,** 105017 (2020).

48. *Star Trek Communicator Online - edical Tricorder* en. https://sites.google.com/site/startrekcommunicatoronline/medical-tricorder.

49. Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. en. *npj Digital Medicine* **3,** 1–10. doi:10.1038/s41746-020-0221-y (2020).

50. Wasylewicz, A. T. & Scheepers-Hoeks, A. Clinical Decision Support Systems. *Fundamentals of clinical data science,* 153–169 (2019).

51. Hak, F., Guimarães, T. & Santos, M. Towards Effective Clinical Decision Support Systems: A Systematic Review. en. *PLoS ONE* **17.** doi:10.1371/journal.pone.0272846 (2022).

52. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big Data in Healthcare: Management, Analysis and Future Prospects. *Journal of Big Data* **6,** 54. doi:10.1186/s40537-019-0217-0 (2019).

53. Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S. & Aggarwal, S. Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy,* 1–20. doi:10.1007/s13132-022-01096-6 (2023).

54. Naithani, N., Sinha, S., Misra, P., Vasudevan, B. & Sahu, R. Precision Medicine: Concept and Tools. *Medical Journal, Armed Forces India* **77,** 249–257. doi:10.1016/j.mjafi.2021.06.021 (2021).

55. Cowie, M. R. *et al.* Electronic Health Records to Facilitate Clinical Research. en. *Clinical Research in Cardiology* **106,** 1–9. doi:10.1007/s00392-016-1025-6 (2017).

56. Poinsignon, T., Poulain, P., Gallopin, M. & Lelandais, G. en. in *Machine Learning for Brain Disorders* (ed Colliot, O.) 313–330 (Springer US, New York, NY, 2023). ISBN: 978-1-07-163195-9. doi:10.1007/978-1-0716-3195-9_10.

57. Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H. & Raad, A. Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review. *Sensors (Basel, Switzerland)* **23,** 828. doi:10.3390/s23020828 (2023).

58. Ip, J. E. Wearable Devices for Cardiac Rhythm Diagnosis and Management. *JAMA* **321,** 337–338. doi:10.1001/jama.2018.20437 (2019).

59. Lu, L., Zhang, J., Xie, Y., Gao, F., Xu, S., Wu, X. & Ye, Z. Wearable Health Devices in Health Care: Narrative Systematic Review. *JMIR mHealth and uHealth* **8,** e18907. doi:10.2196/18907 (2020).

60. Takei, K., Honda, W., Harada, S., Arie, T. & Akita, S. Toward Flexible and Wearable Human-Interactive Health-Monitoring Devices. en. *Advanced Healthcare Materials* **4,** 487–500. doi:10.1002/adhm.201400546 (2015).

61. Mitchell, K. J. What Is Complex about Complex Disorders? *Genome Biology* **13,** 237. doi:10.1186/gb-2012-13-1-237 (2012).

62. Villalobos, P. & Wistuba, I. I. Lung Cancer Biomarkers. *Hematology/oncology clinics of North America* **31,** 13–29. doi:10.1016/j.hoc.2016.08.006 (2017).

63. Alharthi, J. & Eslam, M. Biomarkers of Metabolic (Dysfunction)-Associated Fatty Liver Disease: An Update. *Journal of Clinical and Translational Hepatology* **10,** 134–139. doi:10.14218/JCTH.2021.00248 (2022).

64. Skelly, A. C., Dettori, J. R. & Brodt, E. D. Assessing Bias: The Importance of Considering Confounding. en. *Evidence-Based Spine-Care Journal* **3,** 9. doi:10.1055/s-0031-1298595 (2012).

65. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* en. ISBN: 978-0-262-33737-3 (MIT Press, 2016).

66. Janiesch, C., Zschech, P. & Heinrich, K. Machine Learning and Deep Learning. en. *Electronic Markets* **31,** 685–695. doi:10.1007/s12525-021-00475-2 (2021).

67. Alzubi, J., Nayyar, A. & Kumar, A. *Machine Learning from Theory to Algorithms: An Overview* in *Journal of Physics: Conference Series* **1142** (2018), 012012.

68. Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P. & Ghassemi, M. Deep Reinforcement Learning for Sepsis Treatment. *arXiv preprint arXiv:1711.09602* (2017).

69. Fernando, T., Gammulle, H., Denman, S., Sridharan, S. & Fookes, C. Deep Learning for Medical Anomaly Detection–a Survey. *ACM Computing Surveys (CSUR)* **54,** 1–37 (2021).

70. Kramer, O. en. in *Machine Learning for Evolution Strategies* (ed Kramer, O.) 45–53 (Springer International Publishing, Cham, 2016). ISBN: 978-3-319-33383-0. doi:10.1007/978-3-319-33383-0_5.

71. Lemeshow, S., Klar, J., Teres, D., Avrunin, J. S., Gehlbach, S. H., Rapoport, J. & Rué, M. Mortality Probability Models for Patients in the Intensive Care Unit for 48 or 72 Hours: A Prospective, Multicenter Study. *Critical care medicine* **22,** 1351–1358 (1994).

72. Wagner, D. P. & Draper, E. A. Acute Physiology and Chronic Health Evaluation (APACHE II) and Medicare Reimbursement. *Health care financing review* **1984,** 91 (1984).

73. Le Gall, J.-R., Lemeshow, S. & Saulnier, F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. *Jama* **270,** 2957–2963 (1993).

74. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. doi:10.1023/A:1010933404324 (2001).

75. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. en. *SN Computer Science* **2,** 160. doi:10.1007/s42979-021-00592-x (2021).

76. Zhang, Y.-H., Li, Z., Zeng, T., Pan, X., Chen, L., Liu, D., Li, H., Huang, T. & Cai, Y.-D. Distinguishing Glioblastoma Subtypes by Methylation Signatures. *Frontiers in Genetics* **11** (2020).

77. Dejaegher, J., Solie, L., Hunin, Z., Sciot, R., Capper, D., Siewert, C., Van Cauter, S., Wilms, G., van Loon, J., Ectors, N., Fieuws, S., Pfister, S. M., Van Gool, S. W. & De Vleeschouwer, S. DNA Methylation Based Glioblastoma Subclassification Is Related to Tumoral T-Cell Infiltration and Patient Survival. *Neuro-Oncology* **23,** 240–250. doi:10.1093/neuonc/noaa247 (2020).

78. Capper, D. *et al.* DNA Methylation-Based Classification of Central Nervous System Tumours. en. *Nature* **555,** 469–474. doi:10.1038/nature26000 (2018).

79. *About Glioblastoma* en-US. https://braintumor.org/events/glioblastoma-awareness-day/about-glioblastoma/.

80. Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., *et al.* DNA Methylation-Based Classification of Central Nervous System Tumours. *Nature* **555,** 469–474 (2018).

81. Sturm, D. *et al.* Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. English. *Cancer Cell* **22,** 425–437. doi:10.1016/j.ccr.2012.08.024 (2012).

82. Zhang, Z. A Gentle Introduction to Artificial Neural Networks. *Annals of Translational Medicine* **4,** 370. doi:10.21037/atm.2016.06.20 (2016).

83. Sinaga, K. P. & Yang, M.-S. Unsupervised K-Means Clustering Algorithm. *IEEE access* **8,** 80716–80727 (2020).

84. Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres Terre, H., Shams, Z., Jamnik, M. & Liò, P. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *Frontiers in Genetics* **10** (2019).

85. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).

86. Sohn, K., Lee, H. & Yan, X. Learning Structured Output Representation Using Deep Conditional Generative Models. *Advances in neural information processing systems* **28** (2015).

87. Aguila, A. L., Chapman, J. & Altmann, A. Multi-Modal Variational Autoencoders for Normative Modelling across Multiple Imaging Modalities. *arXiv preprint arXiv:2303.12706* (2023).

88. Talaei Khoei, T., Ould Slimane, H. & Kaabouch, N. Deep Learning: Systematic Review, Models, Challenges, and Research Directions. en. *Neural Computing and Applications* **35,** 23103–23124. doi:10.1007/s00521-023-08957-4 (2023).

89. Kaur, D., Uslu, S., Rittichier, K. J. & Durresi, A. Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys* **55,** 39:1–39:38. doi:10.1145/3491209 (2022).

90. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. & Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Information fusion* **99,** 101805 (2023).

91. Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R. & Díaz-Rodríguez, N. Explainable Artificial Intelligence (Xai) on Timeseries Data: A Survey. *arXiv preprint arXiv:2104.00950* (2021).

92. Caton, S. & Haas, C. Fairness in Machine Learning: A Survey. *ACM Computing Surveys* (2020).

93. Kiritchenko, S. & Mohammad, S. M. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv preprint arXiv:1805.04508* (2018).

94. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. in *Ethics of Data and Analytics* 254–264 (Auerbach Publications, 2022).

95. Du, M., Yang, F., Zou, N. & Hu, X. Fairness in Deep Learning: A Computational Perspective. *IEEE Intelligent Systems* **36,** 25–34 (2020).

96. Meng, C., Trinh, L., Xu, N., Enouen, J. & Liu, Y. Interpretability and Fairness Evaluation of Deep Learning Models on MIMIC-IV Dataset. en. *Scientific Reports* **12,** 7166. doi:10.1038/s41598-022-11012-2 (2022).

97. Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., von Tengg-Kobligk, H., Summers, R. M. & Wiest, R. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. eng. *Radiology. Artificial Intelligence* **2,** e190043. doi:10.1148/ryai.2020190043 (2020).

98. Gastounioti, A. & Kontos, D. Is it time to get rid of black boxes and cultivate trust in AI? *Radiology: Artificial Intelligence* **2,** e200088 (2020).

99. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. English. *The Lancet Digital Health* **3,** e745–e750. doi:10.1016/S2589-7500(21)00208-9 (2021).

100. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. *What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use* in *Machine Learning for Healthcare Conference* (2019), 359–380.

101. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. & Zhong, C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys* **16,** 1–85. doi:10.1214/21-SS133 (2022).

102. Ludwig, J. & Mullainathan, S. *Machine Learning as a Tool for Hypothesis Generation* tech. rep. (National Bureau of Economic Research, 2023).

103. Watson, D. S. & Floridi, L. en. in *Ethics, Governance, and Policies in Artificial Intelligence* (ed Floridi, L.) 185–219 (Springer International Publishing, Cham, 2021). ISBN: 978-3-030-81907-1. doi:10.1007/978-3-030-81907-1_11.

104. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. *Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission* in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), 1721–1730.

105. Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., *et al.* Transfer Learning Enables Predictions in Network Biology. *Nature,* 1–9 (2023).

106. Cannarsa, M. Ethics Guidelines for Trustworthy AI. *The Cambridge handbook of lawyering in the digital age,* 283–297 (2021).

107. Antonini, C. *Navigating the EU AI Act: How Explainable AI Simplifies Regulatory Compliance* 2023. https://positivethinking.tech/insights/navigating-the-eu-ai-act-how-explainable-ai-simplifies-regulatory-compliance/.

108. *Key Issue 5: Transparency Obligations - EU AI Act* https://www.euaiact.com/key-issue/5.

109. Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G. & De Hert, P. Bridging the Gap between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine* **17,** 72–85 (2022).

110. Montanez, C. A. C., Fergus, P., Chalmers, C., Malim, N., Abdulaimma, B., Reilly, D. & Falciani, F. SAERMA: Stacked Autoencoder Rule Mining Algorithm for the Interpretation of Epistatic Interactions in GWAS for Extreme Obesity. English. *IEEE Access* **8,** 112379–112392 (2020).

111. Shams, Z., Dimanov, B., Kola, S., Simidjievski, N. & Terre, H. A. REM: An Integrative Rule Extraction Methodology for Explainable Data Analysis in Healthcare. *medRxiv*. doi:10.1101/2021.01.25.21250459.abstract (2021).

112. Psichogios, D. C. & Ungar, L. H. A Hybrid Neural Network-First Principles Approach to Process Modeling. *AIChE Journal* **38,** 1499–1511 (1992).

113. Ribeiro, M. T., Singh, S. & Guestrin, C. " *Why Should i Trust You?*" *Explaining the Predictions of Any Classifier* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1135–1144.

114. Yang, G., Ye, Q. & Xia, J. Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, Two Showcases and Beyond. *Information Fusion* **77,** 29–52. doi:10.1016/j.inffus.2021.07.016 (2022).

115. Garson, G. D. Interpreting Neural-Network Connection Weights. *AI expert* **6,** 46–51 (1991).

116. Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F. & Zhu, Y. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Front. genet.* **13,** 806842. doi:10.3389/fgene.2022.806842 (2022).

117. Yu, T. AIME: Autoencoder-Based Integrative Multi-Omics Data Embedding That Allows for Confounder Adjustments. *PLoS Computational Biology* **18,** e1009826 (2022).

118. Magnusson, R., Tegner, J. N. & Gustafsson, M. Deep Neural Network Prediction of Genome-Wide Transcriptome Signatures - beyond the Black-Box. *npj syst. biol. appl.* **8,** 9. doi:10.1038/s41540-022-00218-9 (02 23).

119. Dwivedi, S. K., Tjärnberg, A., Tegnér, J. & Gustafsson, M. Deriving Disease Modules from the Compressed Transcriptional Space Embedded in a Deep Autoencoder. en. *Nature Communications* **11,** 856. doi:10.1038/s41467-020-14666-6 (2020).

120. Martínez-Enguita, D., Dwivedi, S. K., Jörnsten, R. & Gustafsson, M. NCAE: Data-Driven Representations Using a Deep Network-Coherent DNA Methylation Autoencoder Identify Robust Disease and Risk Factor Signatures. *Briefings in Bioinformatics* **24,** bbad293. doi:10.1093/bib/bbad293 (2023).

121. Lundberg, S. M. & Lee, S.-I. in *Advances in Neural Information Processing Systems 30* (eds Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R.) 4765–4774 (Curran Associates, Inc., 2017).

122. van Hilten, A., Kushner, S. A., Kayser, M., Ikram, M. A., Adams, H. H. H., Klaver, C. C. W., Niessen, W. J. & Roshchupkin, G. V. GenNet Framework: Interpretable Deep Learning for Predicting Phenotypes from Genetic Data. en. *Communications Biology* **4,** 1–9. doi:10.1038/s42003-021-02622-z (2021).

123. Hao, J., Kim, Y., Kim, T.-K. & Kang, M. PASNet: Pathway-Associated Sparse Deep Neural Network for Prognosis Prediction from High-Throughput Data. en. *BMC Bioinformatics* **19,** 510. doi:10.1186/s12859-018-2500-z (2018).

124. Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., Arnoff, T. E., Park, J., Hahn, W. C. & Van Allen, E. M. Biologically Informed Deep Neural Network for Prostate Cancer Discovery. en. *Nature* **598,** 348–352. doi:10.1038/s41586-021-03922-4 (2021).

125. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA Is an Interpretable Generative Model for Inferring Biological Network Activity in Single-Cell Transcriptomics. en. *Nature Communications* **12,** 5684. doi:10.1038/s41467-021-26017-0 (2021).

126. Madsen, M. B., Skrede, S., Bruun, T., Arnell, P., Rosén, A., Nekludov, M., Karlsson, Y., Bergey, F., Saccenti, E., Martins Dos Santos, V. a. P., Perner, A., Norrby-Teglund, A. & Hyldegaard, O. Necrotizing Soft Tissue Infections - a Multicentre, Prospective Observational Study (INFECT): Protocol and Statistical Analysis Plan. eng. *Acta Anaesthesiologica Scandinavica* **62,** 272–279. doi:10.1111/aas.13024 (2018).

127. Peetermans, M., de Prost, N., Eckmann, C., Norrby-Teglund, A., Skrede, S. & De Waele, J. J. Necrotizing Skin and Soft-Tissue Infections in the Intensive Care Unit. en. *Clinical Microbiology and Infection* **26,** 8–17. doi:10.1016/j.cmi.2019.06.031 (2020).

128. Stevens, D. L. & Bryant, A. E. Necrotizing Soft-Tissue Infections. *New England Journal of Medicine* **377,** 2253–2265. doi:10.1056/NEJMra1600673 (2017).

129. Madsen, M. B. *et al.* Patient's Characteristics and Outcomes in Necrotising Soft-Tissue Infections: Results from a Scandinavian, Multicentre, Prospective Cohort Study. eng. *Intensive Care Medicine* **45,** 1241–1251. doi:10.1007/s00134-019-05730-x (2019).

130. Jabbour, G., El-Menyar, A., Peralta, R., Shaikh, N., Abdelrahman, H., Mudali, I. N., Ellabib, M. & Al-Thani, H. Pattern and Predictors of Mortality in Necrotizing Fasciitis Patients in a Single Tertiary Hospital. *World Journal of Emergency Surgery* **11,** 40. doi:10.1186/s13017-016-0097-y (2016).

131. van Stigt, S. F. L., de Vries, J., Bijker, J. B., Mollen, R. M. H. G., Hekma, E. J., Lemson, S. M. & Tan, E. C. T. H. Review of 58 Patients with Necrotizing Fasciitis in the Netherlands. *World Journal of Emergency Surgery* **11,** 21. doi:10.1186/s13017-016-0080-7 (2016).

132. Audureau, E., Hua, C., de Prost, N., Hemery, F., Decousser, J. W., Bosc, R., Lepeule, R., Chosidow, O. & Sbidian, E. Mortality of Necrotizing Fasciitis: Relative Influence of Individual and Hospital-Level Factors, a Nationwide Multilevel Study, France, 2007–12. en. *British Journal of Dermatology* **177,** 1575–1582. doi:10.1111/bjd.15615 (2017).

133. Tom, L. K., Maine, R. G., Wang, C. S., Parent, B. A., Bulger, E. M. & Keys, K. A. Comparison of Traditional and Skin-Sparing Approaches for Surgical Treatment of Necrotizing Soft-Tissue Infections. eng. *Surgical Infections* **21,** 363–369. doi:10.1089/sur.2019.263 (2020).

134. Al-Qurayshi, Z., Nichols, R. L., Killackey, M. T. & Kandil, E. Mortality Risk in Necrotizing Fasciitis: National Prevalence, Trend, and Burden. eng. *Surgical Infections* **21,** 840–852. doi:10.1089/sur.2019.277 (2020).

135. Horn, D. L., Shen, J., Roberts, E., Wang, T. N., Li, K. S., O'Keefe, G. E., Cuschieri, J., Bulger, E. M. & Robinson, B. R. Predictors of Mortality, Limb Loss, and Discharge Disposition at Admission among Patients with Necrotizing Skin and Soft Tissue Infections. en. *Journal of Trauma and Acute Care Surgery* **89,** 186–191. doi:10.1097/TA.0000000000002636 (2020).

136. Hakkarainen, T. W., Burkette Ikebata, N., Bulger, E. & Evans, H. L. Moving beyond Survival as a Measure of Success: Understanding the Patient Experience of Necrotizing Soft-Tissue Infections. eng. *The Journal of Surgical Research* **192,** 143–149. doi:10.1016/j.jss.2014.05.006 (2014).

137. Fagerdahl, A.-M., Knudsen, V. E., Egerod, I. & Andersson, A. E. Patient Experience of Necrotising Soft-Tissue Infection from Diagnosis to Six Months after Intensive Care Unit Stay: A Qualitative Content Analysis. eng. *Australian Critical Care: Official Journal of the Confederation of Australian Critical Care Nurses* **33,** 187–192. doi:10.1016/j.aucc.2019.02.001 (2020).

138. Knudsen, V. E., Andersson, A. E., Fagerdahl, A.-M. & Egerod, I. Experiences of Family Caregivers the First Six Months after Patient Diagnosis of Necrotising Soft Tissue Infection: A Thematic Analysis. eng. *Intensive & Critical Care Nursing* **49,** 28–36. doi:10.1016/j.iccn.2018.05.005 (2018).

139. Urbina, T., Canoui-Poitrine, F., Hua, C., Layese, R., Alves, A., Ouedraogo, R., Bosc, R., Sbidian, E., Chosidow, O., Dessap, A. M., de Prost, N. & Henri Mondor Hospital Necrotizing Fasciitis Group. Long-Term Quality of Life in Necrotizing Soft-Tissue Infection Survivors: A Monocentric Prospective Cohort Study. eng. *Annals of Intensive Care* **11,** 102. doi:10.1186/s13613-021-00891-9 (2021).

140. Endorf, F. W., Supple, K. G. & Gamelli, R. L. The Evolving Characteristics and Care of Necrotizing Soft-Tissue Infections. eng. *Burns: Journal of the International Society for Burn Injuries* **31,** 269–273. doi:10.1016/j.burns.2004.11.008 (2005).

141. Nawijn, F., Wassenaar, E. C. E., Smeeing, D. P. J., Vlaminckx, B. J. M., Reinders, J. S. K., Wille, J., Leenen, L. P. H. & Hietbrink, F. Exhaustion of the Immune System by Group A Streptococcus Necrotizing Fasciitis: The Occurrence of Late Secondary Infections in a Retrospective Study. *Trauma Surgery & Acute Care Open* **4,** e000272. doi:10.1136/tsaco-2018-000272 (2019).

142. Lauerman, M. H., Scalea, T. M., Eglseder, W. A., Pensy, R., Stein, D. M. & Henry, S. Physiology, Not Modern Operative Approach, Predicts Mortality in Extremity Necrotizing Soft Tissue Infections at a High-Volume Center. eng. *Surgery,* S0039-6060(18)30090–4. doi:10.1016/j.surg.2018.02.013 (2018).

143. Goh, T., Goh, L. G., Ang, C. H. & Wong, C. H. Early Diagnosis of Necrotizing Fasciitis. eng. *The British Journal of Surgery* **101,** e119–125. doi:10.1002/bjs.9371 (2014).

144. Wong, C.-H., Khin, L.-W., Heng, K.-S., Tan, K.-C. & Low, C.-O. The LRINEC (Laboratory Risk Indicator for Necrotizing Fasciitis) Score: A Tool for Distinguishing Necrotizing Fasciitis from Other Soft Tissue Infections. eng. *Critical Care Medicine* **32,** 1535–1541. doi:10.1097/01.ccm.0000129486.35458.7d (2004).

145. Fernando, S. M., Tran, A., Cheng, W., Rochwerg, B., Kyeremanteng, K., Seely, A. J. E., Inaba, K. & Perry, J. J. Necrotizing Soft Tissue Infection: Diagnostic Accuracy of Physical Examination, Imaging, and LRINEC Score: A Systematic Review and Meta-Analysis. eng. *Annals of Surgery* **269,** 58–65. doi:10.1097/SLA.0000000000002774 (2019).

146. Vincent, J. .-., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P. M. & Thijs, L. G. The SOFA (Sepsis-Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure. en. *Intensive Care Medicine* **22,** 707–710. doi:10.1007/BF01709751 (1996).

147. Le Gall, J. R., Lemeshow, S. & Saulnier, F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. eng. *JAMA* **270,** 2957–2963. doi:10.1001/jama.270.24.2957 (1993).

148. Moreno, R. P., Metnitz, P. G. H., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., Le Gall, J.-R. & on behalf of the SAPS 3 Investigators. SAPS 3—From Evaluation of the Patient to Evaluation of the Intensive Care Unit. Part 2: Development of a Prognostic Model for Hospital Mortality at ICU Admission. en. *Intensive Care Medicine* **31,** 1345–1355. doi:10.1007/s00134-005-2763-5 (2005).

149. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: A Severity of Disease Classification System. eng. *Critical Care Medicine* **13,** 818–829 (1985).

150. Song, X., Liu, X., Liu, F. & Wang, C. Comparison of Machine Learning and Logistic Regression Models in Predicting Acute Kidney Injury: A Systematic Review and Meta-Analysis. eng. *International Journal of Medical Informatics* **151,** 104484. doi:10.1016/j.ijmedinf.2021.104484 (2021).

151. Flechet, M., Güiza, F., Schetz, M., Wouters, P., Vanhorebeek, I., Derese, I., Gunst, J., Spriet, I., Casaer, M., den Berghe, G. V. & Meyfroidt, G. AKIpredictor, an Online Prognostic Calculator for Acute Kidney Injury in Adult Critically Ill Patients: Development, Validation and Comparison to Serum Neutrophil Gelatinase-Associated Lipocalin. en. *Intensive Care Medicine* **43.** doi:10.1007/s00134-017-4678-3 (2017).

152. Yee, C. R., Narain, N. R., Akmaev, V. R. & Vemulapalli, V. A Data-Driven Approach to Predicting Septic Shock in the Intensive Care Unit. en. *Biomedical Informatics Insights* **11,** 1178222619885147. doi:10.1177/1178222619885147 (2019).

153. Jahn, M., Rekowski, J., Gerken, G., Kribben, A., Canbay, A. & Katsounas, A. The Predictive Performance of SAPS 2 and SAPS 3 in an Intermediate Care Unit for Internal Medicine at a German University Transplant Center; A Retrospective Analysis. en. *PLOS ONE* **14,** e0222164. doi:10.1371/journal.pone.0222164 (2019).

154. Maslove, D. M. & Wong, H. R. Gene Expression Profiling in Sepsis: Timing, Tissue, and Translational Considerations. *Trends in molecular medicine* **20,** 204–213. doi:10.1016/j.molmed.2014.01.006 (2014).

155. Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J. & Thiesson, B. Explainable Artificial Intelligence Model to Predict Acute Critical Illness from Electronic Health Records. en. *Nature Communications* **11,** 3852. doi:10.1038/s41467-020-17431-x (2020).

156. Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D. & Buchman, T. G. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical care medicine* **46,** 547–553. doi:10.1097/CCM.0000000000002936 (2018).

157. Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. en. *npj Digital Medicine* **3,** 1–10. doi:10.1038/s41746-020-0221-y (2020).

158. Hietbrink, F., Bode, L. G., Riddez, L., Leenen, L. P. H. & van Dijk, M. R. Triple Diagnostics for Early Detection of Ambivalent Necrotizing Fasciitis. eng. *World journal of emergency surgery: WJES* **11,** 51. doi:10.1186/s13017-016-0108-z (2016).

159. Schünemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G. E., Williams, J. W., Kunz, R., Craig, J., Montori, V. M., Bossuyt, P. & Guyatt, G. H. Grading Quality of Evidence and Strength of Recommendations for Diagnostic Tests and Strategies. en. *BMJ* **336,** 1106–1110. doi:10.1136/bmj.39500.677199.AE (2008).

160. Sartelli, M. *et al.* World Society of Emergency Surgery (WSES) Guidelines for Management of Skin and Soft Tissue Infections. *World Journal of Emergency Surgery* **9,** 57. doi:10.1186/1749-7922-9-57 (2014).

161. Stevens, D. L., Bisno, A. L., Chambers, H. F., Dellinger, E. P., Goldstein, E. J. C., Gorbach, S. L., Hirschmann, J. V., Kaplan, S. L., Montoya, J. G. & Wade, J. C. Practice Guidelines for the Diagnosis and Management of Skin and Soft Tissue Infections: 2014 Update by the Infectious Diseases Society of America. en. *Clinical Infectious Diseases* **59,** e10–e52. doi:10.1093/cid/ciu296 (2014).

162. Larry M Baddour, D. L. S. Necrotizing Soft Tissue Infections. *UpToDate* (2021).

163. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. & Cournapeau, D. Scikit-Learn: Machine Learning in Python. en. *MACHINE LEARNING IN PYTHON,* 6.

164. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the **Boruta** Package. en. *Journal of Statistical Software* **36.** doi:10.18637/jss.v036.i11 (2010).

165. Le Gall, J. R., Lemeshow, S., Leleu, G., Klar, J., Huillard, J., Rué, M., Teres, D. & Artigas, A. Customized Probability Models for Early Severe Sepsis in Adult Intensive Care Patients. Intensive Care Unit Scoring Group. eng. *JAMA* **273,** 644–650 (1995).

166. Moreno, R., Vincent, J.-L., Matos, R., Mendonça, A., Cantraine, F., Thijs, L., Takala, J., Sprung, C., Antonelli, M., Bruining, H., Willatts, S. & on behalf of the working group on sepsisrelated problems of. The Use of Maximum SOFA Score to Quantify Organ Dysfunction/Failure in Intensive Care. Results of a Prospective, Multicentre Study. en. *Intensive Care Medicine* **25,** 686–696. doi:10.1007/s001340050931 (1999).

167. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics* **12,** 77. doi:10.1186/1471-2105-12-77 (2011).

168. Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z. & Wang, K. Predicting 30-Days Mortality for MIMIC-III Patients with Sepsis-3: A Machine Learning Approach Using XGboost. *Journal of Translational Medicine* **18,** 462. doi:10.1186/s12967-020-02620-5 (2020).

169. Giannini, H. M., Ginestra, J. C., Chivers, C., Draugelis, M., Hanish, A., Schweickert, W. D., Fuchs, B. D., Meadows, L., Lynch, M., Donnelly, P. J., Pavan, K., Fishman, N. O., Hanson, C. W. I. & Umscheid, C. A. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice*. en-US. *Critical Care Medicine* **47,** 1485–1492. doi:10.1097/CCM.0000000000003891 (2019).

170. Flechet, M., Falini, S., Bonetti, C., Güiza, F., Schetz, M., Van den Berghe, G. & Meyfroidt, G. Machine Learning versus Physicians' Prediction of Acute Kidney Injury in Critically Ill Adults: A Prospective Evaluation of the AKIpredictor. *Critical Care* **23,** 282. doi:10.1186/s13054-019-2563-x (2019).

171. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315,** 801–810. doi:10.1001/jama.2016.0287 (2016).

172. Liu, Z., Meng, Z., Li, Y., Zhao, J., Wu, S., Gou, S. & Wu, H. Prognostic Accuracy of the Serum Lactate Level, the SOFA Score and the qSOFA Score for Mortality among Adults with Sepsis. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* **27,** 51. doi:10.1186/s13049-019-0609-3 (2019).

173. Montassier, E., Batard, E., Segard, J., Hardouin, J.-B., Martinage, A., Le Conte, P. & Potel, G. Base Excess Is an Accurate Predictor of Elevated Lactate in ED Septic Patients. eng. *The American Journal of Emergency Medicine* **30,** 184–187. doi:10.1016/j.ajem.2010.09.033 (2012).

174. Lobo, S. M. A., Lobo, F. R. M., Bota, D. P., Lopes-Ferreira, F., Soliman, H. M., Mélot, C. & Vincent, J.-L. C-Reactive Protein Levels Correlate with Mortality and Organ Failure in Critically Ill Patients. eng. *Chest* **123,** 2043–2049. doi:10.1378/chest.123.6.2043 (2003).

175. Fischer, C. M., Yano, K., Aird, W. C. & Shapiro, N. I. Abnormal Coagulation Tests Obtained in the Emergency Department Are Associated with Mortality in Patients with Suspected Infection. eng. *The Journal of Emergency Medicine* **42,** 127–132. doi:10.1016/j.jemermed.2010.05.007 (2012).

176. Chicco, D. & Oneto, L. Data Analytics and Clinical Feature Ranking of Medical Records of Patients with Sepsis. *BioData Mining* **14,** 12. doi:10.1186/s13040-021-00235-0 (2021).

177. Mao, Q., Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., Shieh, L., Chettipally, U., Fletcher, G., Kerem, Y., Zhou, Y. & Das, R. Multicentre Validation of a Sepsis Prediction Algorithm Using Only Vital Sign Data in the Emergency Department, General Ward and ICU. en. *BMJ Open* **8,** e017833. doi:10.1136/bmjopen-2017-017833 (2018).

178. Antia, R. & Halloran, M. E. Transition to Endemicity: Understanding COVID-19. *Immunity* **54,** 2172–2176. doi:10.1016/j.immuni.2021.09.019 (2021).

179. Kočar, E., Režen, T. & Rozman, D. Cholesterol, Lipoproteins, and COVID-19: Basic Concepts and Clinical Applications. eng. *Biochimica Et Biophysica Acta. Molecular and Cell Biology of Lipids* **1866,** 158849. doi:10.1016/j.bbalip.2020.158849 (2021).

180. Hu, X., Chen, D., Wu, L., He, G. & Ye, W. Low Serum Cholesterol Level among Patients with COVID-19 Infection in Wenzhou, China. *The Lancet*. doi:10.2139/ssrn.3544826 (2020).

181. Wei, C. *et al.* Cholesterol Metabolism–Impact for SARS-CoV-2 Infection Prognosis, Entry, and Antiviral Therapies. *medRxiv*. doi:10.1101/2020.04.16.20068528 (2020).

182. Wei, X., Zeng, W., Su, J., Wan, H., Yu, X., Cao, X., Tan, W. & Wang, H. Hypolipidemia Is Associated with the Severity of COVID-19. *Journal of Clinical Lipidology* **14,** 297–304. doi:10.1016/j.jacl.2020.04.008 (2020).

183. Li, Y., Zhang, Y., Lu, R., Dai, M., Shen, M., Zhang, J. & Cui, Y. Lipid Metabolism Changes in Patients with Severe COVID-19. *Clinica Chimica Acta* **517,** 66–73. doi:10.1016/j.cca.2021.02.011 (2021).

184. Shi, D. *et al.* The Serum Metabolome of COVID-19 Patients Is Distinctive and Predictive. *Metabolism Clinical and Experimental*. doi:10.1016/j.metabol.2021.154739 (2021).

185. Yue, J., Xu, H., Zhou, Y., Liu, W., Han, X., Mao, Q., Li, S., Tam, L. S., Ma, J. & Liu, W. Dyslipidemia Is Related to Mortality in Critical Patients with Coronavirus Disease 2019: A Retrospective Study. *Frontiers in Endocrinology* **12.** doi:10.3389/fendo.2021.611526 (2021).

186. Fabre, B., Fernandez Machulsky, N., Olano, C., Jacobsen, D., Gómez, M. E., Perazzi, B., Zago, V., Zopatti, D., Ferrero, A., Schreier, L. & Berg, G. Remnant Cholesterol Levels Are Associated with Severity and Death in COVID-19 Patients. *Scientific Reports* **12,** 1–6. doi:10.1038/s41598-022-21177-5 (2022).

187. Aladağ, N., Şipal, A., Atabey, R. D., Akbulut, T., Asoğlu, R. & Özdemir, M. Containment Measures Established during the COVID-19 Outbreak and Its Impact on Lipid Profile and Neutrophil to Lymphocyte Ratio. *European Review for Medical and Pharmacological Sciences* **24,** 12510–12515. doi:10.26355/eurrev_202012_24047 (2020).

188. Chen, Y.-M. *et al.* Blood Molecular Markers Associated with COVID-19 Immunopathology and Multi-organ Damage. *The EMBO Journal* **39,** 1–23. doi:10.15252/embj.2020105896 (2020).

189. Hu, X., Chen, D., Wu, L., He, G. & Ye, W. Declined Serum High Density Lipoprotein Cholesterol Is Associated with the Severity of COVID-19 Infection. doi:10.1016/j.cca.2020.07.015 (2020).

190. Tanaka, S., De Tymowski, C., Assadi, M., Zappella, N., Jean-Baptiste, S., Robert, T., Peoch, K., Lortat-Jacob, B., Fontaine, L., Bouzid, D., Tran-Dinh, A., Tashk, P., Meilhac, O. & Montravers, P. Lipoprotein Concentrations over Time in the Intensive Care Unit COVID-19 Patients: Results from the ApoCOVID Study. *PLoS ONE* **15,** 1–15. doi:10.1371/journal.pone.0239573 (2020).

191. Wang, D., Li, R., Wang, J., Jiang, Q., Gao, C., Yang, J., Ge, L. & Hu, Q. Correlation Analysis between Disease Severity and Clinical and Biochemical Characteristics of 143 Cases of COVID-19 in Wuhan, China: A Descriptive Study. *BMC Infectious Diseases* **20,** 1–9. doi:10.1186/s12879-020-05242-w (2020).

192. Wang, G., Zhang, Q., Zhao, X., Dong, H., Wu, C., Wu, F., Yu, B., Lv, J., Zhang, S., Wu, G., Wu, S., Wang, X., Wu, Y. & Zhong, Y. Low High-Density Lipoprotein Level Is Correlated with the Severity of COVID-19 Patients: An Observational Study. *Lipids in Health and Disease* **19,** 1–7. doi:10.1186/s12944-020-01382-9 (2020).

193. Alcántara-Alonso, E., Molinar-Ramos, F., González-López, J. A., Alcántara-Alonso, V., Muñoz-Pérez, M. A., Lozano-Nuevo, J. J., Benítez-Maldonado, D. R. & Mendoza-Portillo, E. High Triglyceride to HDL-Cholesterol Ratio as a Biochemical Marker of Severe Outcomes in COVID-19 Patients. *Clinical Nutrition ESPEN* **44,** 437–444. doi:10.1016/j.clnesp.2021.04.020 (2021).

194. Aparisi, Á., Iglesias-echeverría, C. & Ybarra-falcón, C. Low-Density Lipoprotein Cholesterol Levels Are Associated with Poor Clinical Outcomes in COVID-19. *Nutrition, Metabolism & Cardiovascular Diseases* **6,** 2619–2627. doi:10.1016/j.numecd.2021.06.016 (2021).

195. Lingwood, D. & Simons, K. Lipid Rafts as a Membrane-Organizing Principle. *Science* **327,** 46–50. doi:10.1126/science.1174621 (2010).

196. Simons, K. & Ikonen, E. Functional Rafts in Cell Membranes. en. *Nature* **387,** 569–572. doi:10.1038/42408 (1997).

197. Rezen, T., Rozman, D., Pascussi, J. M. & Monostory, K. Interplay between Cholesterol and Drug Metabolism. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1814,** 146–160. doi:10.1016/j.bbapap.2010.05.014 (2011).

198. Kovač, U., Skubic, C., Bohinc, L., Rozman, D. & Režen, T. Oxysterols and Gastrointestinal Cancers around the Clock. *Frontiers in Endocrinology* **10,** 1–19. doi:10.3389/fendo.2019.00483 (2019).

199. Skubic, C. & Rozman, D. Sterols from the Post-Lanosterol Part of Cholesterol Synthesis: Novel Signaling Players. *Mammalian Sterols,* 1–22. doi:10.1007/978-3-030-39684-8_1 (2020).

200. Rodríguez-Acebes, S., de La Cueva, P., Fernández-Hernando, C., Ferruelo, A. J., Lasunción, M. A., Rawson, R. B., Martínez-Botas, J. & Gómez-Coronado, D. Desmosterol Can Replace Cholesterol in Sustaining Cell Proliferation and Regulating the SREBP Pathway in a Sterol-Δ24-Reductase-Deficient Cell Line. *Biochemical Journal* **420,** 305–318 (2009).

201. Brown, A. J. & Sharpe, L. J. in (eds Ridgway, N. D., Lipoproteins, M. & Membranes (Sixth Edition), R. S. B. T. .-. B. o. L.) 327–358 (Elsevier, Boston, 2016). ISBN: 978-0-444-63438-2. doi:10.1016/B978-0-444-63438-2.00011-0.

202. Kandutsch, A. & Russell, A. Preputial Gland Tumor Sterols. *Journal of Biological Chemistry* **235,** 2256–2261. doi:10.1016/s0021-9258(18)64608-3 (1960).

203. Belič, A., Pompon, D., Monostory, K., Kelly, D., Kelly, S. & Rozman, D. An Algorithm for Rapid Computational Construction of Metabolic Networks: A Cholesterol Biosynthesis Example. *Computers in Biology and Medicine* **43,** 471–480. doi:10.1016/j.compbiomed.2013.02.017 (2013).

204. Liang, W. *et al.* Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients with COVID-19. *JAMA Internal Medicine* **180,** 1081–1089. doi:10.1001/jamainternmed.2020.2033 (2020).

205. Tanaka, S., Couret, D., Tran-Dinh, A., Duranteau, J., Montravers, P., Schwendeman, A. & Meilhac, O. High-Density Lipoproteins during Sepsis: From Bench to Bedside. *Critical Care* **24,** 1–11. doi:10.1186/s13054-020-02860-3 (2020).

206. Cirstea, M., Walley, K. R., Russell, J. A., Brunham, L. R., Genga, K. R. & Boyd, J. H. Decreased High-Density Lipoprotein Cholesterol Level Is an Early Prognostic Marker for Organ Dysfunction and Death in Patients with Suspected Sepsis. *Journal of Critical Care* **38,** 289–294. doi:10.1016/j.jcrc.2016.11.041 (2017).

207. Chien, J.-Y., Jerng, J.-S., Yu, C.-J. & Yang, P.-C. Low Serum Level of High-Density Lipoprotein Cholesterol Is a Poor Prognostic Factor for Severe Sepsis. *Critical Care Medicine* **33.** doi:10.1097/01.ccm.0000171183.79525.6b (2005).

208. Barlage, S., Liebisch, G. & Glu, T. Changes in HDL-Associated Apolipoproteins Relate to Mortality in Human Sepsis and Correlate to Monocyte and Platelet Activation. *Intensive Care Medicine* **35,** 1877–1885. doi:10.1007/s00134-009-1609-y (2009).

209. van Leeuwen, H. J., Heezius, E. C. J. M., Dallinga, G. M., van Strijp, J. A. G., Verhoef, J. & van Kessel, K. P. M. Lipoprotein Metabolism in Patients with Severe Sepsis. English. *Critical care medicine* **31,** 1359–1366. doi:10.1097/01.CCM.0000059724.08290.51 (2003-05, 2003-5).

210. Drobnik, W., Liebisch, G., Audebert, F. X., Fröhlich, D., Glück, T., Vogel, P., Rothe, G. & Schmitz, G. Plasma Ceramide and Lysophosphatidylcholine Inversely Correlate with Mortality in Sepsis Patients. *Journal of Lipid Research* **44,** 754–761. doi:10.1194/jlr.M200401-JLR200 (2003).

211. Lima, W. G., Souza, N. A., Fernandes, S. O. A., Cardoso, V. N. & Godói, I. P. Serum Lipid Profile as a Predictor of Dengue Severity: A Systematic Review and Meta-Analysis. *Reviews in Medical Virology* **29,** 1–13. doi:10.1002/rmv.2056 (2019).

212. Shen, B. *et al.* Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* **182,** 59–72.e15. doi:10.1016/j.cell.2020.05.032 (2020).

213. Masana, L. *et al.* Low HDL and High Triglycerides Predict COVID-19 Severity. *Scientific Reports* **11,** 1–9. doi:10.1038/s41598-021-86747-5 (2021).

214. Barman, A. H., Selcen, A., Dogan, O. & Atıcı, A. Prognostic Significance of Temporal Changes of Lipid Profile in COVID-19 Patients. *Obesity Medicine* **14,** 100373. doi:10.1016/j.obmed.2021.100373 (2021).

215. Caterino, M. *et al.* Dysregulation of Lipid Metabolism and Pathological Inflammation in Patients with COVID-19. *Scientific Reports* **11,** 1–10. doi:10.1038/s41598-021-82426-7 (2021).

216. Dei Cas, M., Ottolenghi, S., Morano, C., Rinaldo, R., Roda, G., Chiumello, D., Centanni, S., Samaja, M. & Paroni, R. Link between Serum Lipid Signature and Prognostic Factors in COVID-19 Patients. en. *Scientific Reports* **11,** 21633. doi:10.1038/s41598-021-00755-z (2021).

217. Sun, J. T., Chen, Z., Nie, P., Ge, H., Shen, L., Yang, F., Qu, X. L., Ying, X. Y., Zhou, Y., Wang, W., Zhang, M. & Pu, J. Lipid Profile Features and Their Associations with Disease Severity and Mortality in Patients with COVID-19. *Frontiers in Cardiovascular Medicine* **7,** 1–12. doi:10.3389/fcvm.2020.584987 (2020).

218. Wang, T., Cao, Y., Zhang, H., Wang, Z., Man, C. H., Yang, Y., Chen, L., Xu, S., Yan, X., Zheng, Q. & Wang, Y. P. COVID-19 Metabolism: Mechanisms and Therapeutic Targets. *MedComm* **3,** 1–24. doi:10.1002/mco2.157 (2022).

219. Masoodi, M., Peschka, M., Schmiedel, S., Haddad, M., Frye, M., Maas, C., Lohse, A., Huber, S., Kirchhof, P., Nofer, J. R. & Renné, T. Disturbed Lipid and Amino Acid Metabolisms in COVID-19 Patients. *Journal of Molecular Medicine* **100,** 555–568. doi:10.1007/s00109-022-02177-4 (2022).

220. Ciccarelli, M. *et al.* Untargeted Lipidomics Reveals Specific Lipid Profiles in COVID-19 Patients with Different Severity from Campania Region (Italy). *Journal of Pharmaceutical and Biomedical Analysis journal* **217.** doi:10.1016/j.jpba.2022.114827 (2022).

221. Aydın, S., Aksakal, E., Aydınyılmaz, F., Gülcü, O., Saraç, İ., Kalkan, K., Aydemir, S., Doğan, R., Aksu, U. & Tanboğa, İ. H. Relationship between Blood Lipid Levels and Mortality in Hospitalized COVID-19 Patients. *Angiology* **73,** 724–733. doi:10.1177/00033197211072346 (2022).

222. Janneh, A. H., Kassir, M. F., Dwyer, C. J., Chakraborty, P., Pierce, J. S., Flume, P. A., Li, H., Nadig, S. N., Mehrotra, S. & Ogretmen, B. Alterations of Lipid Metabolism Provide Serologic Biomarkers for the Detection of Asymptomatic versus Symptomatic COVID-19 Patients. *Scientific Reports* **11,** 1–10. doi:10.1038/s41598-021-93857-7 (2021).

223. Ballout, R. A., Kong, H., Sampson, M., Otvos, J. D., Cox, A. L., Agbor-Enoh, S. & Remaley, A. T. The NIH Lipo-COVID Study: A Pilot NMR Investigation of Lipoprotein Subfractions and Other Metabolites in Patients with Severe COVID-19. *Biomedicines* **9.** doi:10.3390/biomedicines9091090 (2021).

224. Torretta, E., Garziano, M., Poliseno, M., Capitanio, D., Biasin, M., Santantonio, T. A., Clerici, M., Lo Caputo, S., Trabattoni, D. & Gelfi, C. Severity of Covid-19 Patients Predicted by Serum Sphingolipids Signature. *International Journal of Molecular Sciences* **22.** doi:10.3390/ijms221910198 (2021).

225. Bai, Y., Huang, W., Li, Y., Lai, C., Huang, S., Wang, G., He, Y., Hu, L. & Chen, C. Lipidomic Alteration of Plasma in Cured COVID-19 Patients Using Ultra High-Performance Liquid Chromatography with High-Resolution Mass Spectrometry. *Bioscience Reports* **41,** 1–12. doi:10.1042/BSR20204305 (2021).

226. Liu, Y., Pan, Y., Yin, Y., Chen, W. & Li, X. Association of Dyslipidemia with the Severity and Mortality of Coronavirus Disease 2019 (COVID-19): A Meta-Analysis. *Virology Journal* **18,** 1–11. doi:10.1186/s12985-021-01604-1 (2021).

227. Wang, Y. *et al.* Prognostic Value of Leucocyte to High-Density Lipoprotein-Cholesterol Ratios in COVID-19 Patients and the Diabetes Subgroup. *Frontiers in Endocrinology* **12,** 1–10. doi:10.3389/fendo.2021.727419 (2021).

228. Mercorelli, B., Luganini, A., Celegato, M., Palù, G., Gribaudo, G., Lepesheva, G. I. & Loregian, A. The Clinically Approved Antifungal Drug Posaconazole Inhibits Human Cytomegalovirus Replication. **64,** 1–14. doi:10.1128/AAC.00056-20 (2020).

229. Zheng, Y. H., Plemenitas, A., Fielding, C. J. & Peterlin, B. M. Nef Increases the Synthesis of and Transports Cholesterol to Lipid Rafts and HIV-1 Progeny Virions. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 8460–8465. doi:10.1073/pnas.1437453100 (2003).

230. Sheridan, D. A., Shawa, I. T., Thomas, E. L., Felmlee, D. J., Bridge, S. H., Neely, D., Cobbold, J. F., Holmes, E., Bassendine, M. F. & Taylor-Robinson, S. D. Infection with the Hepatitis C Virus Causes Viral Genotype-Specific Differences in Cholesterol Metabolism and Hepatic Steatosis. *Scientific Reports* **12,** 1–11. doi:10.1038/s41598-022-09588-w (2022).

231. Rodgers, M. A., Villareal, V. A., Schaefer, E. A., Peng, L. F., Kathleen, E., Chung, R. T. & Yang, P. L. Lipid Metabolite Profiling Identifies Desmosterol Metabolism as a New Antiviral Target for Hepatitis C Virus. **134,** 6896–6899. doi:10.1021/ja207391q.Lipid (2013).

232. Costello, D. A., Villareal, V. A. & Yang, P. L. Desmosterol Increases Lipid Bilayer Fluidity during Hepatitis C Virus Infection. **2,** 852–862. doi:10.1021/acsinfecdis.6b00086 (2012).

233. Huang, S., Zhou, C., Yuan, Z., Xiao, H. & Wu, X. The Clinical Value of High-Density Lipoprotein in the Evaluation of New Coronavirus Pneumonia. *Advances in Clinical and Experimental Medicine* **30,** 153–156. doi:10.17219/ACEM/130606 (2021).

234. Li, G., Du, L., Cao, X., Wei, X., Jiang, Y., Lin, Y., Nguyen, V., Tan, W. & Wang, H. Follow-up Study on Serum Cholesterol Profiles and Potential Sequelae in Recovered COVID-19 Patients. *BMC Infectious Diseases* **21,** 1–10. doi:10.1186/s12879-021-05984-1 (2021).

235. He, X. *et al.* COVID-19 Induces New-Onset Insulin Resistance and Lipid Metabolic Dysregulation via Regulation of Secreted Metabolic Factors. *Signal Transduction and Targeted Therapy* **6.** doi:10.1038/s41392-021-00822-x (2021).

236. Bizkarguenaga, M. *et al.* Uneven Metabolic and Lipidomic Profiles in Recovered COVID-19 Patients as Investigated by Plasma NMR Metabolomics. *NMR in Biomedicine* **35,** 1–10. doi:10.1002/nbm.4637 (2022).

237. Wu, Q. *et al.* Altered Lipid Metabolism in Recovered SARS Patients Twelve Years after Infection. *Scientific Reports* **7,** 1–12. doi:10.1038/s41598-017-09536-z (2017).

238. Tian, D. & Ye, Q. Hepatic Complications of COVID-19 and Its Treatment. *Journal of Medical Virology* **92,** 1818–1824. doi:10.1002/jmv.26036 (2020).

239. Zhong, P., Xu, J., Yang, D., Shen, Y., Wang, L., Feng, Y., Du, C., Song, Y., Wu, C., Hu, X. & Sun, Y. COVID-19-Associated Gastrointestinal and Liver Injury: Clinical Features and Potential Mechanisms. *Signal Transduction and Targeted Therapy* **5.** doi:10.1038/s41392-020-00373-7 (2020).

240. Saviano, A., Wrensch, F., Ghany, M. G. & Baumert, T. F. Liver Disease and Coronavirus Disease 2019: From Pathogenesis to Clinical Care. *Hepatology* **74,** 1088–1100. doi:10.1002/hep.31684 (2021).

241. Popescu, M., Ştefan, O. M., Ştefan, M., Văleanu, L. & Tomescu, D. ICU-Associated Costs during the Fourth Wave of the COVID-19 Pandemic in a Tertiary Hospital in a Low-Vaccinated Eastern European Country. *International Journal of Environmental Research and Public Health* **19.** doi:10.3390/ijerph19031781 (2022).

242. Skevaki, C., Fragkou, P. C., Cheng, C., Xie, M. & Renz, H. Laboratory Characteristics of Patients Infected with the Novel SARS-CoV-2 Virus. *Journal of Infection* **81,** 205–212. doi:10.1016/j.jinf.2020.06.039 (2020).

243. Hentsch, L., Cocetta, S., Allali, G., Santana, I., Eason, R., Adam, E. & Janssens, J. P. Breathlessness and COVID-19: A Call for Research. *Respiration* **100,** 1016–1026. doi:10.1159/000517400 (2021).

244. Zhou, F. *et al.* Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China: A Retrospective Cohort Study. *The Lancet* **395,** 1054–1062. doi:10.1016/S0140-6736(20)30566-3 (2020).

245. Jin Zhang, J., Dong, X., yuan Cao, Y., dong Yuan, Y., Yang, b. Y., qin Yan, Y., Akdis, C. A. & dong Gao, Y. Clinical Characteristics of 140 Patients Infected with SARS-CoV-2 in Wuhan, China. *Allergy: European Journal of Allergy and Clinical Immunology* **75,** 1730–1741. doi:10.1111/all.14238 (2020).

246. Borghesi, A., Zigliani, A., Golemi, S., Carapella, N., Maculotti, P., Farina, D. & Maroldi, R. Chest X-Ray Severity Index as a Predictor of in-Hospital Mortality in Coronavirus Disease 2019: A Study of 302 Patients from Italy. *International Journal of Infectious Diseases* **96,** 291–293. doi:10.1016/j.ijid.2020.05.021 (2020).

247. Rubin, G. D. *et al.* The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic. *Chest* **158,** 106–16. doi:10.1016/j.chest.2020.04.003 (2020).

248. Meher, G., Bhattacharjya, S. & Chakraborty, H. Membrane Cholesterol Modulates Oligomeric Status and Peptide-Membrane Interaction of Severe Acute Respiratory Syndrome Coronavirus Fusion Peptide. *Journal of Physical Chemistry B* **123,** 10654–10662. doi:10.1021/acs.jpcb.9b08455 (2019).

249. Abu-Farha, M., Thanaraj, T. A., Qaddoumi, M. G., Hashem, A., Abubaker, J. & Al-Mulla, F. The Role of Lipid Metabolism in COVID-19 Virus Infection and as a Drug Target. *International Journal of Molecular Sciences* **21.** doi:10.3390/ijms21103544 (2020).

250. Salimi, H., Johnson, J., Flores, M. G., Zhang, M. S., O'Malley, Y., Houtman, J. C., Schlievert, P. M. & Haim, H. The Lipid Membrane of HIV-1 Stabilizes the Viral Envelope Glycoproteins and Modulates Their Sensitivity to Antibody Neutralization. *Journal of Biological Chemistry* **295,** 348–362. doi:10.1074/jbc.RA119.009481 (2020).

251. Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C. & Huang, L. Comparing Different Machine Learning Techniques for Predicting COVID-19 Severity. *Infectious Diseases of Poverty* **11,** 1–9. doi:10.1186/s40249-022-00946-4 (2022).

252. Zhou, K. *et al.* Eleven Routine Clinical Features Predict COVID-19 Severity Uncovered by Machine Learning of Longitudinal Measurements. *Computational and Structural Biotechnology Journal* **19,** 3640–3649. doi:10.1016/j.csbj.2021.06.022 (2021).

253. Statsenko, Y., Al Zahmi, F., Habuza, T., Gorkom, V. K. N. & Zaki, N. Prediction of COVID-19 Severity Using Laboratory Findings on Admission: Informative Values, Thresholds, ML Model Performance. *BMJ Open* **11.** doi:10.1136/bmjopen-2020-044500 (2021).

254. Yan, L. *et al.* An Interpretable Mortality Prediction Model for COVID-19 Patients. *Nature Machine Intelligence* **2,** 283–288. doi:10.1038/s42256-020-0180-7 (2020).

255. Moreno-Pérez, Ó., Andrés, M., León-Ramirez, J. M., Sánchez-Payá, J., Boix, V., Gil, J. & Merino, E. The COVID-GRAM Tool for Patients Hospitalized with COVID-19 in Europe. *JAMA Internal Medicine* **181,** 1000. doi:10.1001/jamainternmed.2021.0491 (2021-07-01, 2021-7-1).

256. Sebastian, A. *et al.* The Usefulness of the COVID-GRAM Score in Predicting the Outcomes of Study Population with COVID-19. *International Journal of Environmental Research and Public Health* **19.** doi:10.3390/ijerph191912537 (2022).

257. Nardo, A. D., Schneeweiss-Gleixner, M., Bakail, M., Dixon, E. D., Lax, S. F. & Trauner, M. Pathophysiological Mechanisms of Liver Injury in COVID-19. *Liver International* **41,** 20–32. doi:10.1111/liv.14730 (2021).

258. Patel, K. P., Patel, P. A., Vunnam, R. R., Hewlett, A. T., Jain, R., Jing, R. & Vunnam, S. R. Gastrointestinal, Hepatobiliary, and Pancreatic Manifestations of COVID-19. *Journal of Clinical Virology* **128,** 104386 (2020).

259. Kolesova, O., Vanaga, I., Laivacuma, S., Derovs, A., Kolesovs, A., Radzina, M., Platkajis, A., Eglite, J., Hagina, E., Arutjunana, S., Putrins, D. S., Storozenko, J., Rozentale, B. & Viksna, L. Intriguing Findings of Liver Fibrosis Following COVID-19. *BMC Gastroenterology* **21,** 4–12. doi:10.1186/s12876-021-01939-7 (2021).

260. Aby, E. S. *et al.* Long-Term Clinical Outcomes of Patients with COVID-19 and Chronic Liver Disease: US Multicenter COLD Study. *Hepatology Communications* **7,** e8874–e8874. doi:10.1097/01.hc9.0000897224.68874.de (2023).

261. Barbara, J. M., Gatt, J., Xuereb, R. A., Tabone Adami, N., Darmanin, J., Erasmi, R., G Xuereb, R., Barbara, C., Stephen, F. & Jane Magri, C. Clinical Outcomes at Medium-Term Follow-up of COVID-19. *Journal of the Royal College of Physicians of Edinburgh* **52,** 220–227. doi:10.1177/14782715221124617 (2022).

262. Lu, J. Y., Ho, S. L., Buczek, A., Fleysher, R., Hou, W., Chacko, K. & Duong, T. Q. Clinical Predictors of Recovery of COVID-19 Associated-Abnormal Liver Function Test 2 Months after Hospital Discharge. *Scientific Reports* **12,** 1–10. doi:10.1038/s41598-022-22741-9 (2022).

263. COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19) Treatment Guidelines. National Institutes of Health. Available at https://www.covid19treatmentgu Accessed [4/3/2023].

264. Skubic, C., Vovk, I., Rozman, D. & Križman, M. Simplified LC-MS Method for Analysis of Sterols in Biological Samples. *Molecules* **25.** doi:10.3390/molecules25184116 (2020).

265. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

266. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **36,** 1–13. doi:10.18637/jss.v036.i11 (2010).

267. Breiman, L. Random Forest. *Machine Learning* **45,** 5–32. doi:10.1023/A:1010933404324 (2001).

268. Rasmussen, C. E. Gaussian Processes in Machine Learning. *Lecture Notes in Computer Science* **3176,** 67–75. doi:10.1007/978-3-540-28650-9_4 (2004).

269. Thongkam, J., Xu, G. & Zhang, Y. AdaBoost Algorithm with Random Forests for Predicting Breast Cancer Survivability. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),* 3062–3069. doi:10.1109/IJCNN.2008.4634231 (2008).

270. Dreiseitl, S. & Ohno-Machado, L. Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. *Journal of Biomedical Informatics* **35,** 352–359. doi:10.1016/S1532-0464(03)00034-0 (2002).

271. Kramer, O. in, 13–23 (2013). ISBN: 978-3-642-38652-7. doi:10.1007/978-3-642-38652-7_2.

272. Gardner, M. W. & Dorling, S. R. Artificial Neural Networks (the Multilayer Perceptron)—a Review of Applications in the Atmospheric Sciences. *Atmospheric Environment* **32,** 2627–2636. doi:10.1016/S1352-2310(97)00447-0 (1998).

273. Rish, I. An Empirical Study of the Naive Bayes Classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence,* 41–46 (2001).

274. Tharwat, A. Linear vs. Quadratic Discriminant Analysis Classifier: A Tutorial. *International Journal of Applied Pattern Recognition* **3,** 145. doi:10.1504/ijapr.2016.079050 (2016).

275. Sartelli, M. *et al.* 2018 WSES/SIS-E Consensus Conference: Recommendations for the Management of Skin and Soft-Tissue Infections. eng. *World journal of emergency surgery: WJES* **13,** 58. doi:10.1186/s13017-018-0219-9 (2018).

276. Linnér, A., Darenberg, J., Sjölin, J., Henriques-Normark, B. & Norrby-Teglund, A. Clinical Efficacy of Polyspecific Intravenous Immunoglobulin Therapy in Patients with Streptococcal Toxic Shock Syndrome: A Comparative Observational Study. eng. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **59,** 851–857. doi:10.1093/cid/ciu449 (2014).

277. Bruun, T., Rath, E., Madsen, M. B., Oppegaard, O., Nekludov, M., Arnell, P., Karlsson, Y., Babbar, A., Bergey, F., Itzek, A., Hyldegaard, O., Norrby-Teglund, A., Skrede, S. & INFECT Study Group. Risk Factors and Predictors of Mortality in Streptococcal Necrotizing Soft-Tissue Infections: A Multicenter Prospective Study. eng. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **72,** 293–300. doi:10.1093/cid/ciaa027 (2021).

278. Parks, T., Wilson, C., Curtis, N., Norrby-Teglund, A. & Sriskandan, S. Polyspecific Intravenous Immunoglobulin in Clindamycin-Treated Patients With Streptococcal Toxic Shock Syndrome: A Systematic Review and Meta-Analysis. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **67,** 1434–1436. doi:10.1093/cid/ciy401 (2018).

279. Niessen, F. A., de Jong, V. M., Janssen, S. & Boel, C. H. E. Dutch guideline on necrotizing soft tissue infections. dut. *Nederlands Tijdschrift Voor Geneeskunde* **164,** D4737 (2020).

280. Johannesen, T. B. *et al.* Increase in Invasive Group A Streptococcal Infections and Emergence of Novel, Rapidly Expanding Sub-Lineage of the Virulent Streptococcus Pyogenes M1 Clone, Denmark, 2023. en. *Eurosurveillance* **28,** 2300291. doi:10.2807/1560-7917.ES.2023.28.26.2300291 (2023).

281. Abo, Y.-N., Oliver, J., McMinn, A., Osowicki, J., Baker, C., Clark, J. E., Blyth, C. C., Francis, J. R., Carr, J., Smeesters, P. R., Crawford, N. W. & Steer, A. C. Increase in Invasive Group A Streptococcal Disease among Australian Children Coinciding with Northern Hemisphere Surges. English. *The Lancet Regional Health – Western Pacific* **41.** doi:10.1016/j.lanwpc.2023.100873 (2023).

282. Aboulhosn, A., Sanson, M. A., Vega, L. A., Segura, M. G., Joseph, M., McNeil, J. C. & Flores, A. R. Increases in Group A Streptococcal Infections in the Pediatric Population in Houston, Texas, 2022. *Clinical Infectious Diseases* **77,** 351–354. doi:10.1093/cid/ciad197 (2023).

283. Suijker, J., Pijpe, A., Hoogerbrug, D., Heymans, M. W., van Zuijlen, P. P. M., Halm, J. A., Meij-de Vries, A. & NSTI Knowledge Collaborative Group. IDENTIFICATION OF POTENTIALLY MODIFIABLE FACTORS TO IMPROVE RECOGNITION AND OUT-

COME OF NECROTIZING SOFT-TISSUE INFECTIONS. eng. *Shock (Augusta, Ga.)* **61,** 585–591. doi:10.1097/SHK.0000000000002325 (2024).

284. Katz, S., Suijker, J., Hardt, C., Madsen, M. B., Vries, A. M.-d., Pijpe, A., Skrede, S., Hyldegaard, O., Solligård, E., Norrby-Teglund, A., Saccenti, E. & Martins dos Santos, V. A. P. Decision Support System and Outcome Prediction in a Cohort of Patients with Necrotizing Soft-Tissue Infections. *International Journal of Medical Informatics* **167,** 104878. doi:10.1016/j.ijmedinf.2022.104878 (2022).

285. Cox, D. R. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* **20,** 215–232. doi:10.1111/j.2517-6161.1958.tb00292.x (1958).

286. Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* (MIT press Cambridge, MA, 2006).

287. Youssef, A., Pencina, M., Thakur, A., Zhu, T., Clifton, D. & Shah, N. H. External Validation of AI Models in Health Should Be Replaced with Recurring Local Validation. en. *Nature Medicine* **29,** 2686–2687. doi:10.1038/s41591-023-02540-z (2023).

288. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. en, 10.

289. Willis, R. N., Guidry, C. A., Horn, C. B., Gilsdorf, D., Davies, S. W., Dietch, Z. C. & Sawyer, R. G. Predictors of Monomicrobial Necrotizing Soft Tissue Infections. eng. *Surgical Infections* **16,** 533–537. doi:10.1089/sur.2014.189 (2015).

290. Suijker, J., Hofmans, F. A. C., van Zuijlen, P. P. M., Cense, H. A., Bonjer, H. J. & Vries, A. M.-d. Approaches to Surgical Debridement in Necrotizing Soft Tissue Infections: Outcomes of an Animated, Interactive Survey. eng. *World Journal of Surgery* **46,** 1051–1058. doi:10.1007/s00268-022-06470-8 (2022).

291. Movert, E., Bolarin, J. S., Valfridsson, C., Velarde, J., Skrede, S., Nekludov, M., Hyldegaard, O., Arnell, P., Svensson, M., Norrby-Teglund, A., Cho, K. H., Elhaik, E., Wessels, M. R., Råberg, L. & Carlsson, F. Interplay between Human STING Genotype and Bacterial NADase Activity Regulates Inter-Individual Disease Variability. en. *Nature Communications* **14,** 4008. doi:10.1038/s41467-023-39771-0 (2023).

292. Mitchell, W. G., Dee, E. C. & Celi, L. A. Generalisability through Local Validation: Overcoming Barriers Due to Data Disparity in Healthcare. *BMC Ophthalmology* **21,** 228. doi:10.1186/s12886-021-01992-6 (2021).

293. Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres-Terre, H., Shams, Z., Jamnik, M. & Liò, P. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *bioRxiv,* 719542. doi:10.1101/719542 (2019).

294. Goh, W. W. B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. eng. *Trends in Biotechnology* **35,** 498–507. doi:10.1016/j.tibtech.2017.02.012 (2017).

295. Čuklina, J., Pedrioli, P. G. A. & Aebersold, R. en. in *Mass Spectrometry Data Analysis in Proteomics* (ed Matthiesen, R.) 373–387 (Springer, New York, NY, 2020). ISBN: 978-1-4939-9744-2. doi:10.1007/978-1-4939-9744-2_16.

296. Pourhoseingholi, M., Baghestani, A. & Vahedi, M. How to control confounding effects by statistical analysis. en. *Gastroenterol Hepatol Bed Bench* **5,** 79–83 (2012).

297. Radhakrishnan, A., Friedman, S. F., Khurshid, S., Ng, K., Batra, P., Lubitz, S. A., Philippakis, A. A. & Uhler, C. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nature Communications* **14,** 2436 (2023).

298. Lawry Aguila, A., Chapman, J., Janahi, M. & Altmann, A. *Conditional VAEs for Confound Removal and Normative Modelling of Neurodegenerative Diseases* en. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (eds Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) (Springer Nature Switzerland, Cham, 2022), 430–440. ISBN: 978-3-031-16431-6. doi:10.1007/978-3-031-16431-6_41.

299. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial Deconfounding Autoencoder for Learning Robust Gene Expression Embeddings. *Bioinformatics* **36,** i573–i582. doi:10.1093/bioinformatics/btaa796 (2020).

300. Bahrami, M., Maitra, M., Nagy, C., Turecki, G., Rabiee, H. R. & Li, Y. Deep Feature Extraction of Single-Cell Transcriptomes by Generative Adversarial Network. *Bioinformatics* **37,** 1345–1351. doi:10.1093/bioinformatics/btaa976 (2021).

301. Liu, X., Li, B., Bron, E., Niessen, W., Wolvius, E. & Roshchupkin, G. en. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science(* (ed Bruijne, M.) (Springer, 2021). doi:10.1007/978-3-030-87240-3_78. https://doi.org/10.1007/978-3-030-87240-3_78.

302. de Lima Camillo, L. P., Lapierre, L. R. & Singh, R. A Pan-Tissue DNA-Methylation Epigenetic Clock Based on Deep Learning. en. *npj Aging* **8,** 1–15. doi:10.1038/s41514-022-00085-y (2022).

303. Tu, H., Wen, C. P., Tsai, S. P., Chow, W.-H., Wen, C., Ye, Y., Zhao, H., Tsai, M. K., Huang, M., Dinney, C. P., Tsao, C. K. & Wu, X. Cancer risk associated with chronic diseases and disease markers: prospective cohort study. eng. *BMJ (Clinical research ed.)* **360,** k134. doi:10.1136/bmj.k134 (2018).

304. Kartsonaki, C. *et al.* Circulating proteins and risk of pancreatic cancer: a case-subcohort study among Chinese adults. eng. *International Journal of Epidemiology* **51,** 817–829. doi:10.1093/ije/dyab274 (2022).

305. Odegaard, A. O., Koh, W. P., Yu, M. C. & Yuan, J. M. Body mass index and risk of colorectal cancer in Chinese Singaporeans: the Singapore Chinese Health Study. *Cancer* **117,** 3841–3849 (2011).

306. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics* **45,** 1113–1120. doi:10.1038/ng.2764 (2013).

307. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G. & Noushmehr, H. TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data. *Nucleic Acids Research* **44,** e71. doi:10.1093/nar/gkv1507 (2016).

308. Chen, F., Zhang, Y., Bossé, D., Lalani, A.-K. A., Hakimi, A. A., Hsieh, J. J., Choueiri, T. K., Gibbons, D. L., Ittmann, M. & Creighton, C. J. Pan-urologic cancer genomic subtypes that transcend tissue of origin. eng. *Nature Communications* **8,** 199. doi:10.1038/s41467-017-00289-x (2017).

309. Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173,** 291–304 (2018).

310. Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2,** 401–404 (2012).

311. Sohn, K., Lee, H. & Yan, X. *Learning Structured Output Representation Using Deep Conditional Generative Models* in *Advances in Neural Information Processing Systems* (eds Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R.) **28** (Curran Associates, Inc., 2015).

312. Fan, Y. J. Autoencoder Node Saliency: Selecting Relevant Latent Representations. *Pattern Recognition* **88,** 643–653. doi:10.1016/j.patcog.2018.12.015 (2019).

313. Kiselev, V., Kirschner, K., Schaub, M., Andrews, T., Yiu, A., Chandra, T., Natarajan, K., Reik, W., Barahona, M., Green, A. & Hemberg, M. SC3: consensus clustering of single-cell RNA-seq data. nl. *Nat Methods.* Epub 2017 Mar 27. PMID: 28346451; PMCID: PMC5410170. doi:10.1038/nmeth.4236. (2017-05-14).

314. Falcon, W. & The PyTorch Lightning team. *PyTorch Lightning* version 1.4. 2019. doi:10.5281/zenodo.3828935. https://github.com/Lightning-AI/lightning.

315. Kamat, A. M., Hahn, N. M., Efstathiou, J. A., Lerner, S. P., Malmström, P.-U., Choi, W., Guo, C. C., Lotan, Y. & Kassouf, W. Bladder cancer. *The Lancet* **388,** 2796–2810 (2016).

316. Horne, T. K. & Cronje, M. J. Cancer Tissue Classification, Associated Therapeutic Implications and PDT as an Alternative. en. *Anticancer Research* **37,** 2785–2807 (2017).

317. Uyar, B., Ronen, J., Franke, V., Gargiulo, G. & Akalin, A. Multi-omics and deep learning provide a multifaceted view of cancer. *bioRxiv,* 2021–09 (2021).

318. González-Reymúndez, A. & Vázquez, A. I. Multi-Omic Signatures Identify Pan-Cancer Classes of Tumors beyond Tissue of Origin. en. *Scientific Reports* **10,** 8341. doi:10.1038/s41598-020-65119-5 (2020).

319. Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C. & Guo, Y. *Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification* in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2019), 765–769. doi:10.1109/BIBM47256.2019.8983228.

320. Wang, X., Zhou, R., Zhao, K., Leow, A., Zhang, Y. & He, L. *Normative Modeling Via Conditional Variational Autoencoder and Adversarial Learning to Identify Brain Dysfunction in Alzheimer's Disease* in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023), 1–4. doi:10.1109/ISBI53787.2023.10230377.

321. Yu, T. AIME: Autoencoder-Based Integrative Multi-Omics Data Embedding That Allows for Confounder Adjustments. en. *PLOS Computational Biology* **18,** e1009826. doi:10.1371/journal.pcbi.1009826 (2022).

322. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Fei-Fei, L., Niebles, J. C. & Pohl, K. M. *Representation learning with statistical independence to mitigate bias* in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), 2513–2523.

323. Eltager, M., Abdelaal, T., Charrout, M., Mahfouz, A., Reinders, M. J. T. & Makrodimitris, S. Benchmarking Variational AutoEncoders on Cancer Transcriptomics Data. en. *PLOS ONE* (2023).

324. Owens, A., McInerney, C. & Prise, K. Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (Hepatocellular carcinoma. en. *BMC Bioinformatics* **22,** 563. doi:10.1186/s12859-021-04454-4. https://doi.org/10.1186/s12859-021-04454-4 (2021).

325. Muhammad, H. en. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science(* (ed Shen, D.) (Springer, 2019). doi:10.1007/978-3-030-32239-7_67. https://doi.org/10.1007/978-3-030-32239-7_67.

326. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F. & Rodrigues, F. A. Clustering algorithms: A comparative approach. eng. *PloS One* **14,** e0210236. doi:10.1371/journal.pone.0210236 (2019).

327. Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., Via, A. & Colombo, T. AI applications in functional genomics. *Computational and Structural Biotechnology Journal* **19,** 5762–5790 (2021).

328. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169,** 1177–1186 (2017).

329. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. & Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis* **42,** 60–88 (2017).

330. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **30** (2017).

331. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

332. *Fundamental Texts On European Private Law* (eds Radley-Gardner, O., Beale, H. & Zimmermann, R.) ISBN: 978-1-78225-864-3 978-1-78225-865-0 978-1-78225-866-7 978-1-78225-867-4. doi:10.5040/9781782258674. http://www.bloomsburycollections.com/book/fundamental-texts-on-european-private-law-1 (Hart Publishing, 2016).

333. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics* **24,** 125–137 (2023).

334. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the black box: interpretable machine learning for geneticists. *Trends in genetics* **36,** 442–455 (2020).

335. Zhang, Y., Tino, P., Leonardis, A. & Tang, K. A Survey on Neural Network Interpretability. English. *IEEE Trans. Emerg. Top. Comput. Intell.* **5,** 726–742 (2021).

336. Watson, D. S. Interpretable machine learning for genomics. *Human genetics* **141,** 1499–1513 (2022).

337. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC bioinformatics* **24,** 1–31 (2023).

338. Min, X., Zeng, W., Chen, N., Chen, T. & Jiang, R. Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics* **33,** i92–i101 (2017).

339. Karim, M. R., Cochez, M., Beyan, O., Decker, S. & Lange, C. OncoNetExplainer: Explainable Predictions of Cancer Types Based on Gene Expression Data. *2019 Ieee 19th International Conference on Bioinformatics and Bioengineering,* 415–422 (2019).

340. Karim, M. R., Rahman, A., Jares, J. B., Decker, S. & Beyan, O. A Snapshot Neural Ensemble Method for Cancer-Type Prediction Based on Copy Number Variations. *Neural Comput. Appl.* **32,** 15281–15299 (2019).

341. Lombardo, E., Hess, J., Kurz, C., Riboldi, M., Marschner, S., Baumeister, P., Lauber, K., Pflugradt, U., Walch, A., Canis, M., Klauschen, F., Zitzelsberger, H., Belka, C., Landry, G. & Unger, K. DeepClassPathway: Molecular Pathway Aware Classification Using Explainable Deep Learning. *Eur J Cancer* **176,** 41–49. doi:10.1016/j.ejca.2022.08.033. http://dx.doi.org/10.1016/j.ejca.2022.08.033%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=36191385%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med22&AN=36191385 (0011).

342. Pampari, A., Shcherbina, A., Nair, S., Schreiber, J., Patel, A., Wang, A., Kundu, S., Shrikumar, A. & Kundaje, A. *Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants.* version 0.1.1. 2023. doi:10.5281/zenodo.7567627. https://github.com/kundajelab/chrombpnet.

343. Choi, S. R. & Lee, M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* **12,** 1033 (2023).

344. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521,** 436–444 (2015).

345. Lu, Z., Pu, H., Wang, F., Hu, Z. & Wang, L. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems* **30** (2017).

346. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural networks* **2,** 359–366 (1989).

347. Sanchez-Lengeling, B., Reif, E., Pearce, A. & Wiltschko, A. B. A gentle introduction to graph neural networks. *Distill* **6,** e33 (2021).

348. Michael, K. Y., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J. & Ideker, T. Visible machine learning for biomedicine. *Cell* **173,** 1562–1565 (2018).

349. Abdullah, M., Madain, A. & Jararweh, Y. *ChatGPT: Fundamentals, applications and social impacts* in *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (2022), 1–8.

350. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F. & Wang, B. scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. *bioRxiv,* 2023–04 (2023).

351. Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Song, L. & Zhang, X. Large Scale Foundation Model on Single-cell Transcriptomics. *bioRxiv,* 2023–05 (2023).

352. Watson, D. S. *Interpretable Machine Learning for Genomics* (Springer, 2022).

353. Wohlin, C. *Guidelines for snowballing in systematic literature studies and a replication in software engineering* in *Proceedings of the 18th international conference on evaluation and assessment in software engineering* (2014), 1–10.

354. Kassani, P. H., Lu, F., Le Guen, Y., Belloy, M. E. & He, Z. Deep neural networks with controlled variable selection for the identification of putative causal genetic variants. *Nature Machine Intelligence* **4,** 761–771 (2022).

355. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610,** 704–712 (2022).

356. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., *et al.* The future of digital health with federated learning. *NPJ digital medicine* **3,** 119 (2020).

357. Roth, H. R., Cheng, Y., Wen, Y., Yang, I., Xu, Z., Hsieh, Y.-T., Kersten, K., Harouni, A., Zhao, C., Lu, K., *et al.* Nvidia flare: Federated learning from simulation to real-world. *arXiv preprint arXiv:2210.13291* (2022).

358. Tonner, P. D., Pressman, A. & Ross, D. Interpretable Modeling of Genotype-Phenotype Landscapes with State-of-the-Art Predictive Power. *Proc Natl Acad Sci U S A* **119,** e2114021119. doi:10.1073/pnas.2114021119. http://dx.doi.org/10.1073/pnas.2114021119%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35733251%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=35733251 (06 28).

359. Wang, Y. & Chen, L. DeepPerVar: A Multimodal Deep Learning Framework for Functional Interpretation of Genetic Variants in Personal Genome. *bioRxiv*. doi:10.1101/2022.04.10.487809. https://www.embase.com/search/results?subaction=viewrecord&id=L2018216562&from=export%20http://dx.doi.org/10.1101/2022.04.10.487809 (2022).

360. Demetci, P., Cheng, W., Darnell, G., Zhou, X., Ramachandran, S. & Crawford, L. Multi-Scale Inference of Genetic Trait Architecture Using Biologically Annotated Neural Networks. *PLoS Genet* **17,** e1009754. doi:10.1371/journal.pgen.1009754. http://dx.doi.org/10.1371/journal.pgen.1009754%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34411094%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=34411094 (2021).

361. Hu, J., Yu, W., Dai, Y., Liu, C., Wang, Y. & Wu, Q. A Deep Neural Network for Gastric Cancer Prognosis Prediction Based on Biological Information Pathways. *J. Oncol.* **2022,** 2965166. doi:10.1155/2022/2965166. http://dx.doi.org/10.1155/2022/2965166%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=36117847%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=36117847 (2022).

362. Feng, J., Zhang, H. & Li, F. Investigating the Relevance of Major Signaling Pathways in Cancer Survival Using a Biologically Meaningful Deep Learning Model. *BMC Bioinformatics* **22,** 47. doi:10.1186/s12859-020-03850-6. http://dx.doi.org/10.1186/s12859-020-03850-6%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=

33546587%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med19&AN=33546587 (2021).

363. Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F. & Zhu, Y. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Front. genet.* **13,** 806842. doi:10.3389/fgene.2022.806842. http://dx.doi.org/10.3389/fgene.2022.806842%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35186034%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=35186034 (2022).

364. Arango, G., Kipkogei, E., Jacob, E., Kagiampakis, I. & Patra, A. Explainable Transformer-Based Neural Network Forthe Prediction of Survival Outcomes in Non-SmallCell Lung Cancer (NSCLC). *medRxiv.* doi:10.1101/2021.10.11.21264761.abstract. https://www.medrxiv.org/content/10.1101/2021.10.11.21264761.abstract (2021).

365. Kobayashi, K., Bolatkan, A., Shiina, S. & Hamamoto, R. Fully-Connected Neural Networks with Reduced Parameterization for Predicting Histological Types of Lung Cancer from Somatic Mutations. doi:10.3390/biom10091249. http://dx.doi.org/10.3390/biom10091249%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=32872133%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med18&AN=32872133 (08 28).

366. Schulte-Sasse, R., Budach, S., Hnisz, D. & Marsico, A. Integration of Multiomics Data with Graph Convolutional Networks to Identify New Cancer Genes and Their Associated Molecular Mechanisms. *Nat. Mach. Intell.* **3,** 513–+ (2021).

367. Ghafouri-Fard, S., Taheri, M., Omrani, M. D., Daaee, A. & Mohammad-Rahimi, H. Application of Artificial Neural Network for Prediction of Risk of Multiple Sclerosis Based on Single Nucleotide Polymorphism Genotypes. *J Mol Neurosci* **70,** 1081–1087. doi:10.1007/s12031-020-01514-x. http://dx.doi.org/10.1007/s12031-020-01514-x%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=32152937%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med17&AN=32152937 (2020).

368. Nguyen, N. D., Huang, J. & Wang, D. A Deep Manifold-Regularized Learning Model for Improving Phenotype Prediction from Multi-Modal Data. *Nat Comput Sci* **2,** 38–46. doi:10.1038/s43588-021-00185-x. http://dx.doi.org/10.1038/s43588-021-00185-x%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_

fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35480297%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=35480297 (2022).

369. Van Hilten, A., Kushner, S. A., Kayser, M., Ikram, M. A., Adams, H. H. H., Klaver, C. C. W., Niessen, W. J. & Roshchupkin, G. V. GenNet Framework: Interpretable Deep Learning for Predicting Phenotypes from Genetic Data. *Commun Biol* **4,** 1094. doi:10.1038/s42003-021-02622-z. http://dx.doi.org/10.1038/s42003-021-02622-z%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34535759%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=34535759 (09 17).

370. Raimondi, D., Simm, J., Arany, A., Fariselli, P., Cleynen, I. & Moreau, Y. An Interpretable Low-Complexity Machine Learning Framework for Robust Exome-Based in-Silico Diagnosis of Crohn's Disease Patients. *NAR genom. bioinform.* **2,** lqaa011. doi:10.1093/nargab/lqaa011. http://dx.doi.org/10.1093/nargab/lqaa011%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33575557%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm5&AN=33575557 (2020).

371. Battey, C. J., Coffing, G. C. & Kern, A. D. Visualizing Population Structure with Variational Autoencoders. doi:10.1093/g3journal/jkaa036. http://dx.doi.org/10.1093/g3journal/jkaa036%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33561250%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med19&AN=33561250 (01 18).

372. Ausmees, K. & Nettelblad, C. A Deep Learning Framework for Characterization of Genotype Data. doi:10.1093/g3journal/jkac020. http://dx.doi.org/10.1093/g3journal/jkac020%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35078229%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=35078229 (03 04).

373. Motsinger-Reif, A. A., Reif, D. M., Fanelli, T. J. & Ritchie, M. D. A Comparison of Analytical Methods for Genetic Association Studies. *Genet Epidemiol* **32,** 767–778. doi:10.1002/gepi.20345. http://dx.doi.org/10.1002/gepi.20345%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=

PubMed&dopt=Citation&list_uids=18561203%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med7&AN=18561203 (2008).

374. Badre, A. & Pan, C. LINA: A Linearizing Neural Network Architecture for Accurate First-Order and Second-Order Interpretations. *IEEE Access* **10,** 36166–36176. doi:10.1109/access.2022.3163257. http://dx.doi.org/10.1109/access.2022.3163257%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35462722%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=35462722 (2022).

375. Greenside, P., Shimko, T., Fordyce, P. & Kundaje, A. Discovering Epistatic Feature Interactions from Neural Network Models of Regulatory DNA Sequences. *Bioinformatics* **34,** i629–i637. doi:10.1093/bioinformatics/bty575. http://dx.doi.org/10.1093/bioinformatics/bty575%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=30423062%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med15&AN=30423062 (09 01).

376. Lee, S., Lim, S., Lee, T., Sung, I. & Kim, S. Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics* **36,** 3818–3824 (2020).

377. Yuan, L., Lai, J., Zhao, J., Sun, T., Hu, C., Ye, L., Yu, G. & Yang, Z. Path-ATT-CNN: A Novel Deep Neural Network Method for Key Pathway Identification of Lung Cancer. *Front. genet.* **13,** 896884. doi:10.3389/fgene.2022.896884. http://dx.doi.org/10.3389/fgene.2022.896884%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35783280%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=35783280 (2022).

378. Ma, T. & Zhang, A. Incorporating biological knowledge with factor graph neural network for interpretable deep learning. *arXiv preprint arXiv:1906.00537* (2019).

379. Cho, H. J., Shu, M., Bekiranov, S., Zang, C. & Zhang, A. Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. *Bioinformatics* **39,** btad113 (2023).

380. Zhang, T. H., Hasib, M. M., Chiu, Y. C., Han, Z. F., Jin, Y. F., Flores, M., Chen, Y. & Huang, Y. Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. doi:10.3390/cancers14194763. http://dx.doi.org/10.3390/cancers14194763%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=

PubMed&dopt=Citation&list_uids=36230685%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm7&AN=36230685 (2022).

381. Ramirez, R., Chiu, Y. C., Zhang, S., Ramirez, J., Chen, Y., Huang, Y. & Jin, Y. F. Prediction and Interpretation of Cancer Survival Using Graph Convolution Neural Networks. *Methods* **192,** 120–130. doi:10.1016/j.ymeth.2021.01.004. http://dx.doi.org/10.1016/j.ymeth.2021.01.004%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33484826%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=33484826 (0008).

382. Bourgeais, V., Zehraoui, F., Ben Hamdoune, M. & Hanczar, B. Deep GONet: Self-Explainable Deep Neural Network Based on Gene Ontology for Phenotype Prediction from Gene Expression Data. *BMC Bioinformatics* **22,** 455. doi:10.1186/s12859-021-04370-7. http://dx.doi.org/10.1186/s12859-021-04370-7%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34551707%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med19&AN=34551707 (Suppl 10 2021).

383. Choi, Y., Li, R. & Quon, G. Interpretable Deep Generative Models for Genomics. *bioRxiv*. doi:10.1101/2021.09.15.460498. https://www.embase.com/search/results?subaction=viewrecord&id=L2015042099&from=export%20http://dx.doi.org/10.1101/2021.09.15.460498 (2022).

384. Alessandri, L., Cordero, F., Beccuti, M., Licheri, N., Arigoni, M., Olivero, M., Di Renzo, M. F., Sapino, A. & Calogero, R. Sparsely-Connected Autoencoder (SCA) for Single Cell RNAseq Data Mining. *NPJ Systems Biology and Applications* **7,** 1. doi:10.1038/s41540-020-00162-6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7785742/ (2021).

385. Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-Zadeh, S., Talavera-López, C., Misharin, A. V. & Theis, F. J. Biologically Informed Deep Learning to Infer Gene Program Activity in Single Cells. *bioRxiv*. doi:10.1101/2022.02.05.479217. https://www.embase.com/search/results?subaction=viewrecord&id=L2017266840&from=export%20http://dx.doi.org/10.1101/2022.02.05.479217 (2022).

386. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA Is an Interpretable Generative Model for Inferring Biological Network Activity in Single-Cell Transcriptomics. *Nat Commun* **12,** 5684. doi:10.1038/s41467-021-26017-0. http://dx.doi.org/10.1038/s41467-021-26017-0%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=

34584103%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med19&AN=34584103 (2021).

387. Lemsara, A., Ouadfel, S. & Fröhlich, H. PathME: Pathway Based Multi-Modal Sparse Autoencoders for Clustering of Patient-Level Multi-Omics Data. *BMC Bioinformatics* **21,** 146. doi:10.1186/s12859-020-3465-2 (2020).

388. Pan, X., Burgman, B., Sahni, N. & Yi, S. S. Deep learning based on multi-omics integration identifies potential therapeutic targets in breast cancer. *bioRxiv,* 2022–01 (2022).

389. Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., Arnoff, T. E., Park, J., Hahn, W. C. & Van Allen, E. M. Biologically Informed Deep Neural Network for Prostate Cancer Discovery. *Nature* **598,** 348–352. doi:10.1038/s41586-021-03922-4. http://dx.doi.org/10.1038/s41586-021-03922-4%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34552244%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med19&AN=34552244 (2021).

390. Azher, Z. L., Vaickus, L. J., Salas, L. A., Christensen, B. C. & Levy, J. J. Development of Biologically Interpretable Multimodal Deep Learning Model for Cancer Prognosis Prediction. *bioRxiv*. doi:10.1101/2021.10.30.466610. https://www.embase.com/search/results?subaction=viewrecord&id=L2015896116&from=export%20http://dx.doi.org/10.1101/2021.10.30.466610 (2021).

391. Kaczmarek, E., Jamzad, A., Imtiaz, T., Nanayakkara, J., Renwick, N. & Mousavi, P. Multi-Omic Graph Transformers for Cancer Classification and Interpretation. *Pac Symp Biocomput* **27,** 373–384. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34890164%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=34890164 (2022).

392. Levy, J. J., Chen, Y., Azizgolshani, N., Petersen, C. L., Titus, A. J., Moen, E. L., Vaickus, L. J., Salas, L. A. & Christensen, B. C. MethylSPWNet and MethylCapsNet: Biologically Motivated Organization of DNAm Neural Networks, Inspired by Capsule Networks. *npj Systems Biology and Applications* **7,** 1–16. doi:10.1038/s41540-021-00193-7. https://www.nature.com/articles/s41540-021-00193-7 (1 2021).

393. Cai, Z., Poulos, R. C., Aref, A., Robinson, P. J., Reddel, R. R. & Zhong, Q. Transformer-based deep learning integrates multi-omic data with cancer pathways. *bioRxiv,* 2022–10 (2022).

394. Zhou, M., Zhang, H., Bai, Z., Mann-Krzisnik, D., Wang, F. & Li, Y. Single-cell multi-omic topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures. *bioRxiv,* 2023–01 (2023).

395. Huang, Z., Wang, J., Yan, Z. & Guo, M. Differentially Expressed Genes Prediction by Multiple Self-Attention on Epigenetics Data. doi:10.1093/bib/bbac117. http://dx.doi.org/10.1093/bib/bbac117%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35380603%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=35380603 (05 13).

396. Fiosina, J., Fiosins, M. & Bonn, S. Explainable Deep Learning for Augmentation of Small RNA Expression Profiles. *Journal of Computational Biology*. doi:10.1089/cmb.2019.0320. https://www.liebertpub.com/doi/abs/10.1089/cmb.2019.0320 (2020).

397. Liu, G. & Bichindaritz, I. An Explainable Deep Network Framework with Case-based Reasoning Strategies for Survival Analysis in Cancer (2022).

398. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

399. Xing, X., Yang, F., Li, H., Zhang, J., Zhao, Y., Gao, M., Huang, J. & Yao, J. Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. *Bioinformatics* **38,** 2178–2186 (2022).

400. Yingtaweesittikul, H. & Suphavilai, C. Network-Guided Supervised Learning on Gene Expression Using a Graph Convolutional Neural Network. *bioRxiv*. doi:10.1101/2021.12.27.474240. https://www.embase.com/search/results?subaction=viewrecord&id=L2016438702&from=export%20http://dx.doi.org/10.1101/2021.12.27.474240 (2021).

401. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications* **12,** 5261 (2021).

402. Minoura, K., Abe, K., Nam, H., Nishikawa, H. & Shimamura, T. A Mixture-of-Experts Deep Generative Model for Integrated Analysis of Single-Cell Multiomics Data. *Cell Rep Methods* **1,** 100071. doi:10.1016/j.crmeth.2021.100071. http://dx.doi.org/10.1016/j.crmeth.2021.100071%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35474667%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=35474667 (2021).

403. Janizek, J. D., Spiro, A., Celik, S., Blue, B. W., Russell, J. C., Lee, T. I., Kaeberlin, M. & Lee, S. I. Principled Feature Attribution for Unsupervised Gene Expression Analysis. *bioRxiv*. doi:10.1101/2022.05.03.490535. https://www.embase.com/search/results?subaction=viewrecord&id=L2018760577&from=export%20http://dx.doi.org/10.1101/2022.05.03.490535 (2022).

404. Wang, H., Wu, Z. & Xing, E. P. Removing Confounding Factors Associated Weights in Deep Neural Networks Improves the Prediction Accuracy for Healthcare Applications. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **24,** 54–65 (2019).

405. Zhao, Q., Adeli, E. & Pohl, K. M. Training Confounder-Free Deep Learning Models for Medical Applications. *Nature Communications* **11,** 6010. doi:10.1038/s41467-020-19784-9. https://www.nature.com/articles/s41467-020-19784-9 (1 2020).

406. Holzscheck, N., Falckenhayn, C., Sohle, J., Kristof, B., Siegner, R., Werner, A., Schossow, J., Jurgens, C., Volzke, H., Wenck, H., Winnefeld, M., Gronniger, E. & Kaderali, L. Modeling Transcriptomic Age Using Knowledge-Primed Artificial Neural Networks. *npj aging mech. dis.* **7,** 15. doi:10.1038/s41514-021-00068-5. http://dx.doi.org/10.1038/s41514-021-00068-5%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34075044%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm5&AN=34075044 (2021).

407. Albaradei, S., Albaradei, A., Alsaedi, A., Uludag, M., Thafar, M. A., Gojobori, T., Essack, M. & Gao, X. MetastaSite: Predicting Metastasis to Different Sites Using Deep Learning with Gene Expression Data. *Front. mol. biosci.* **9,** 913602. doi:10.3389/fmolb.2022.913602. http://dx.doi.org/10.3389/fmolb.2022.913602%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35936793%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm&AN=35936793 (2022).

408. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., *et al.* Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).

409. Ribeiro, M. T., Singh, S. & Guestrin, C. "*Why Should I Trust You?*": *Explaining the Predictions of Any Classifier* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (2016), 1135–1144.

410. Meudec, R. *tf-explain* version 0.3.1. 2021. doi:10.5281/zenodo.5711704. https://github.com/sicara/tf-explain.

411. Shrikumar, A., Greenside, P. & Kundaje, A. *Learning important features through propagating activation differences* in *International conference on machine learning* (2017), 3145–3153.

412. Neagu, C.-D., Avouris, N., Kalapanidas, E. & Palade, V. Neural and neuro-fuzzy integration in a knowledge-based system for air quality prediction. *Applied Intelligence* **17,** 141–169 (2002).

413. Pal, N. R., Sharma, A. & Sanadhya, S. K. Deriving meaningful rules from gene expression data for classification. *Journal of Intelligent & Fuzzy Systems* **19,** 171–180 (2008).

414. Chen, C.-F., Feng, X. & Szeto, J. Identification of critical genes in microarray experiments by a Neuro-Fuzzy approach. *Computational Biology and Chemistry* **30,** 372–381 (2006).

415. Jha, A., Quesnel-Vallieres, M., Wang, D., Thomas-Tikhonenko, A., Lynch, K. W. & Barash, Y. Identifying Common Transcriptome Signatures of Cancer by Interpreting Deep Learning Models. *Genome Biol* **23,** 117. doi:10.1186/s13059-022-02681-3. http://dx.doi.org/10.1186/s13059-022-02681-3%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35581644%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=35581644 (2022).

416. Withnell, E., Zhang, X., Sun, K. & Guo, Y. XOmiVAE: An Interpretable Deep Learning Model for Cancer Classification Using High-Dimensional Omics Data. *Briefings in Bioinformatics.* doi:10.1093/bib/bbab315 (11 05).

417. Sun, T., Wei, Y., Chen, W. & Ding, Y. Genome-Wide Association Study-Based Deep Learning for Survival Prediction. *Stat Med* **39,** 4605–4620. doi:10.1002/sim.8743. http://dx.doi.org/10.1002/sim.8743%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=32974946%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med18&AN=32974946 (2022).

418. *AIME: Autoencoder-Based Integrative Multi-Omics Data Embedding That Allows for Confounder Adjustments | PLOS Computational Biology* https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009826.

419. Camillo, L. P. D., Lapierre, L. R. & Singh, R. A Pan-Tissue DNA-Methylation Epigenetic Clock Based on Deep Learning. *npj Aging* **8.** doi:10.1038/s41514-022-00085-y (2022).

420. Van den Broeck, G., Lykov, A., Schleich, M. & Suciu, D. On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research* **74,** 851–886 (2022).

421. Jin, J., Yu, Y., Wang, R., Zeng, X., Pang, C., Jiang, Y., Li, Z., Dai, Y., Su, R., Zou, Q., Nakai, K. & Wei, L. iDNA-ABF: Multi-Scale Deep Biological Language Learning Model for the Interpretable Prediction of DNA Methylations. *Genome Biol* **23,** 219. doi:10.1186/s13059-022-02780-1. http://dx.doi.org/10.1186/s13059-022-02780-1%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=36253864%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med22&AN=36253864 (10 17).

422. Liu, G., Zeng, H. & Gifford, D. K. Visualizing Complex Feature Interactions and Feature Sharing in Genomic Deep Neural Networks. *BMC Bioinformatics* **20,** 401. doi:10.1186/s12859-019-2957-4. http://dx.doi.org/10.1186/s12859-019-2957-4%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=31324140%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med16&AN=31324140 (2019).

423. Dwivedi, K., Rajpal, A., Rajpal, S. & Agarwal, M. An Explainable AI-Driven Biomarker Discovery Framework for Non-Small Cell Lung Cancer Classification. *Computers in Biology ...* https://www.sciencedirect.com/science/article/pii/S0010482523000094 (2023).

424. Chatzianastasis, M., Vazirgiannis, M. & Zhang, Z. Explainable Multilayer Graph Neural Network for Cancer Gene Prediction. *arXiv preprint arXiv ...* https://arxiv.org/abs/2301.08831 (2023).

425. Real, K. S. D. & Rubio, A. Discovering the Mechanism of Action of Drugs with a Novel Sparse Explainable Network. *Available at SSRN 4364890*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4364890.

426. Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F., Kramer, F., Leha, A. & Beisbarth, T. Explaining Decisions of Graph Convolutional Neural Networks: Patient-Specific Molecular Subnetworks Responsible for Metastasis Prediction in Breast Cancer. *Genome Med* **13,** 42. doi:10.1186/s13073-021-00845-7. http://dx.doi.org/10.1186/s13073-021-00845-7%20http://www.ncbi.nlm.nih.

gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33706810%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=33706810 (03 11).

427. Mieth, B., Rozier, A., Rodriguez, J. A., Hohne, M. M. C., Gornitz, N. & Muller, K. R. DeepCOMBI: Explainable Artificial Intelligence for the Analysis and Discovery in Genome-Wide Association Studies. *NAR genom. bioinform.* **3,** lqab065. doi:10.1093/nargab/lqab065. http://dx.doi.org/10.1093/nargab/lqab065%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=34296082%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=pmnm5&AN=34296082 (2021).

428. Yap, M., Johnston, R. L., Foley, H., MacDonald, S., Kondrashova, O., Tran, K. A., Nones, K., Koufariotis, L. T., Bean, C., Pearson, J. V., Trzaskowski, M. & Waddell, N. Verifying Explainability of a Deep Learning Tissue Classifier Trained on RNA-Seq Data. *Sci. rep.* **11,** 2641. doi:10.1038/s41598-021-81773-9. http://dx.doi.org/10.1038/s41598-021-81773-9%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33514769%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med18&AN=33514769 (01 29).

429. Benkirane, H., Pradat, Y., Michiels, S. & Cournede, P. H. CustOmics: A Versatile Deep-Learning Based Strategy for Multi-Omics Integration. *PLoS Comput Biol* **19,** e1010921. doi:10.1371/journal.pcbi.1010921. http://dx.doi.org/10.1371/journal.pcbi.1010921%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=36877736%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=mesx&AN=36877736 (2023).

430. Van de Leur, R. R., Bos, M. N., Taha, K., Sammani, A., Yeung, M. W., van Duijvenboden, S., Lambiase, P. D., Hassink, R. J., van der Harst, P., Doevendans, P. A., Gupta, D. K. & van Es, R. Improving Explainability of Deep Neural Network-Based Electrocardiogram Interpretation Using Variational Auto-Encoders. *European Heart Journal - Digital Health* **3,** 390–404. doi:10.1093/ehjdh/ztac038. https://doi.org/10.1093/ehjdh/ztac038 (2022).

431. Liu, L., Meng, Q., Weng, C., Lu, Q., Wang, T. & Wen, Y. Explainable Deep Transfer Learning Model for Disease Risk Prediction Using High-Dimensional Genomic Data. *PLoS Comput Biol* **18,** e1010328. doi:10.1371/journal.pcbi.1010328. http://dx.doi.org/10.1371/journal.pcbi.1010328%20http://www.ncbi.nlm.nih.gov/entrez/

query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=35839250%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=35839250 (0007).

432. Watson, M., Hasan, B. A. S. & Moubayed, N. A. *Using Model Explanations to Guide Deep Learning Models towards Consistent Explanations for EHR Data* (nature.com, 2022).

433. Jin, Y., Ren, Z., Wang, W. J., Zhang, Y. L., Zhou, L., Yao, X. F. & Wu, T. Classification of Alzheimer's Disease Using Robust TabNet Neural Networks on Genetic Data. *Math. Biosci. Eng.* **20,** 8358–8374. doi:10.3934/mbe.2023366 (2023).

434. Sun, T., Wei, Y., Chen, W. & Ding, Y. Genome-wide association study-based deep learning for survival prediction. *Statistics in medicine* **39,** 4605–4620 (2020).

435. Martens, K. & Yau, C. BasisVAE: Translation-Invariant Feature-Level Clustering with Variational Autoencoders. *International Conference on Artificial Intelligence and Statistics, Vol 108* **108,** 2928–2936 (2020).

436. Liang, B., Gong, H., Lu, L. & Xu, J. Risk Stratification and Pathway Analysis Based on Graph Neural Network and Interpretable Algorithm. *BMC Bioinformatics* **23,** 394. doi:10.1186/s12859-022-04950-1. http://dx.doi.org/10.1186/s12859-022-04950-1%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=36167504%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med22&AN=36167504 (2022).

437. Ghafouri-Fard, S., Taheri, M., Omrani, M. D., Daaee, A., Mohammad-Rahimi, H. & Kazazi, H. Application of Single-Nucleotide Polymorphisms in the Diagnosis of Autism Spectrum Disorders: A Preliminary Study with Artificial Neural Networks. *J Mol Neurosci* **68,** 515–521. doi:10.1007/s12031-019-01311-1. http://dx.doi.org/10.1007/s12031-019-01311-1%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=30937628%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med16&AN=30937628 (2019).

438. Qin, R., Mahal, L. K. & Bojar, D. Deep Learning Explains the Biology of Branched Glycans from Single-Cell Sequencing Data. *bioRxiv*. doi:10.1101/2022.06.27.497708. https://www.embase.com/search/results?subaction=viewrecord&id=L2019456871&from=export%20http://dx.doi.org/10.1101/2022.06.27.497708 (2022).

439. Keyl, P., Bischoff, P. & Dernbach, G. Single-Cell Gene Regulatory Network Prediction by Explainable AI. *Nucleic Acids ...* doi:10.1093/nar/gkac1212/6984592. https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkac1212/6984592 (2023).

440. Jin, T., Nguyen, N. D., Talos, F. & Wang, D. ECMarker: Interpretable Machine Learning Model Identifies Gene Expression Biomarkers Predicting Clinical Outcomes and Reveals Molecular Mechanisms of Human Disease in Early Stages. *Bioinformatics* **37,** 1115–1124. doi:10.1093/bioinformatics/btaa935. http://dx.doi.org/10.1093/bioinformatics/btaa935%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33305308%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med18&AN=33305308 (2021).

441. Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F. K., Rodig, S. J., Lindeman, N. I. & Mahmood, F. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Trans Med Imaging* **41,** 757–770. doi:10.1109/tmi.2020.3021387. http://dx.doi.org/10.1109/tmi.2020.3021387%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=32881682%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med20&AN=32881682 (2020).

442. Søgaard, A. Shortcomings of Interpretability Taxonomies for Deep Neural Networks. *Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI)* (2022).

443. Van Hilten, A., van Rooij, J., consortium, B., Ikram, M. A., Niessen, W. J., van Meurs, J. B. & Roshchupkin, G. V. Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data. *bioRxiv,* 2023–04 (2023).

444. Esser-Skala, W. & Fortelny, N. Reliable interpretability of biology-inspired deep neural networks. *NPJ Systems Biology and Applications* **9,** 50 (2023).

445. Urbanowicz, R. J., Kiralis, J., Sinnott-Armstrong, N. A., Heberling, T., Fisher, J. M. & Moore, J. H. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData mining* **5,** 1–14 (2012).

446. Blumenthal, D. B., Viola, L., List, M., Baumbach, J., Tieri, P. & Kacprowski, T. EpiGEN: an epistasis simulation pipeline. *Bioinformatics* **36,** 4957–4959 (2020).

447. Yang, W. & Gu, C. C. A Whole-Genome Simulator Capable of Modeling High-Order Epistasis for Complex Disease. *Genetic epidemiology* **37,** 686–694 (2013).

448. Jain, S. & Wallace, B. C. *Attention Is Not Explanation* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* NAACL-HLT 2019 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), 3543–3556. doi:10.18653/v1/N19-1357. https://aclanthology.org/N19-1357.

449. Bastings, J. & Filippova, K. *The Elephant in the Interpretability Room: Why Use Attention as Explanation When We Have Saliency Methods?* in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (Association for Computational Linguistics, Online, 2020), 149–155. doi:10.18653/v1/2020.blackboxnlp-1.14. https://www.aclweb.org/anthology/2020.blackboxnlp-1.14.

450. Li, Y., Wu, F.-X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics* **19,** 325–340 (2018).

451. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., Wilson, K. T., Landman, B. & Huo, Y. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering* (2023).

452. Stahlschmidt, S. R., Ulfenborg, B. & Synnergren, J. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics* **23,** bbab569 (2022).

453. Smith, Z. D. & Meissner, A. DNA Methylation: Roles in Mammalian Development. en. *Nature Reviews Genetics* **14,** 204–220. doi:10.1038/nrg3354 (2013).

454. Martin, E. M. & Fry, R. C. Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. *Annual Review of Public Health* **39,** 309–333. doi:10.1146/annurev-publhealth-040617-014629 (2018).

455. Houtepen, L. C. *et al.* Genome-Wide DNA Methylation Levels and Altered Cortisol Stress Reactivity Following Childhood Trauma in Humans. en. *Nature Communications* **7,** 10967. doi:10.1038/ncomms10967 (2016).

456. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. eng. *Circulation. Cardiovascular Genetics* **9,** 436–447. doi:10.1161/CIRCGENETICS.116.001506 (2016).

457. Maas, S. C. E. *et al.* Validated Inference of Smoking Habits from Blood with a Finite DNA Methylation Marker Set. eng. *European Journal of Epidemiology* **34,** 1055–1074. doi:10.1007/s10654-019-00555-w (2019).

458. Seale, K., Horvath, S., Teschendorff, A., Eynon, N. & Voisin, S. Making Sense of the Ageing Methylome. en. *Nature Reviews Genetics* **23,** 585–605. doi:10.1038/s41576-022-00477-6 (2022).

459. Mulder, R. H. *et al.* Epigenome-Wide Change and Variation in DNA Methylation in Childhood: Trajectories from Birth to Late Adolescence. eng. *Human Molecular Genetics* **30,** 119–134. doi:10.1093/hmg/ddaa280 (2021).

460. Campagna, M. P., Xavier, A., Lechner-Scott, J., Maltby, V., Scott, R. J., Butzkueven, H., Jokubaitis, V. G. & Lea, R. A. Epigenome-Wide Association Studies: Current Knowledge, Strategies and Recommendations. *Clinical Epigenetics* **13,** 214. doi:10.1186/s13148-021-01200-8 (2021).

461. Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., Li, R., Xia, L., Zhang, T., Niu, G., Bao, Y. & Zhang, Z. EWAS Atlas: A Curated Knowledgebase of Epigenome-Wide Association Studies. eng. *Nucleic Acids Research* **47,** D983–D988. doi:10.1093/nar/gky1027 (2019).

462. Wang, Z., Wu, X. & Wang, Y. A Framework for Analyzing DNA Methylation Data from Illumina Infinium HumanMethylation450 BeadChip. *BMC Bioinformatics* **19,** 115. doi:10.1186/s12859-018-2096-3 (2018).

463. Laird, P. W. Principles and Challenges of Genome-Wide DNA Methylation Analysis. en. *Nature Reviews Genetics* **11,** 191–203. doi:10.1038/nrg2732 (2010).

464. de Lima Camillo, L. P., Lapierre, L. R. & Singh, R. A Pan-Tissue DNA-Methylation Epigenetic Clock Based on Deep Learning. en. *npj Aging* **8,** 1–15. doi:10.1038/s41514-022-00085-y (2022).

465. Way, G. P. & Greene, C. S. Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. eng. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **23,** 80–91 (2018).

466. Titus, A. J., Wilkins, O. M., Bobak, C. A. & Christensen, B. C. *Unsupervised Deep Learning with Variational Autoencoders Applied to Breast Tumor Genome-Wide DNA Methylation Data with Biologic Feature Extraction* en. Preprint (Bioinformatics, 2018). doi:10.1101/433763.

467. Owens, A. R., McInerney, C. E., Prise, K. M., McArt, D. G. & Jurek-Loughrey, A. Novel Deep Learning-Based Solution for Identification of Prognostic Subgroups in Liver Cancer (Hepatocellular Carcinoma). *BMC Bioinformatics* **22,** 563. doi:10.1186/s12859-021-04454-4 (2021).

468. Qiu, Y. L., Zheng, H. & Gevaert, O. Genomic Data Imputation with Variational Auto-Encoders. English. *GigaScience* **9.** doi:10.1093/gigascience/giaa082 (2020).

469. Choi, J. & Chae, H. methCancer-Gen: A DNA Methylome Dataset Generator for User-Specified Cancer Type Based on Conditional Variational Autoencoder. *BMC Bioinformatics* **21,** 181. doi:10.1186/s12859-020-3516-8 (2020).

470. Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A. & Christensen, B. C. MethylNet: An Automated and Modular Deep Learning Approach for DNA Methylation Analysis. *BMC Bioinformatics* **21,** 108. doi:10.1186/s12859-020-3443-8 (2020).

471. Macías-García, L., Martínez-Ballesteros, M., Luna-Romera, J. M., García-Heredia, J. M., García-Gutiérrez, J. & Riquelme-Santos, J. C. Autoencoded DNA Methylation Data to Predict Breast Cancer Recurrence: Machine Learning Models and Gene-Weight Significance. en. *Artificial Intelligence in Medicine* **110,** 101976. doi:10.1016/j.artmed.2020.101976 (2020).

472. Eltager, M., Abdelaal, T., Charrout, M., Mahfouz, A., Reinders, M. J. & Makrodimitris, S. *Benchmarking Variational AutoEncoders on Cancer Transcriptomics Data* en. Preprint (Bioinformatics, 2023). doi:10.1101/2023.02.09.527832.

473. Hu, Q. & Greene, C. S. Parameter Tuning Is a Key Part of Dimensionality Reduction via Deep Variational Autoencoders for Single Cell RNA Transcriptomics. eng. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **24,** 362–373 (2019).

474. Battey, C. J., Coffing, G. C. & Kern, A. D. Visualizing Population Structure with Variational Autoencoders. en. *G3 Genes|Genomes|Genetics* **11** (ed Sethuraman, A.) jkaa036. doi:10.1093/g3journal/jkaa036 (2021).

475. Choi, Y., Li, R. & Quon, G. *Interpretable Deep Generative Models for Genomics* en. Preprint (Genomics, 2021). doi:10.1101/2021.09.15.460498.

476. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B. & Shen, R. High Density DNA Methylation Array with Single CpG Site Resolution. en. *Genomics. New Genomic Technologies and Applications* **98,** 288–295. doi:10.1016/j.ygeno.2011.07.007 (2011).

477. Shu, H. & Zhu, H. Sensitivity Analysis of Deep Neural Networks. en. *Proceedings of the AAAI Conference on Artificial Intelligence* **33,** 4943–4950. doi:10.1609/aaai.v33i01.33014943 (2019).

478. Magnusson, R., Tegnér, J. N. & Gustafsson, M. Deep Neural Network Prediction of Genome-Wide Transcriptome Signatures – beyond the Black-Box. en. *npj Systems Biology and Applications* **8,** 1–8. doi:10.1038/s41540-022-00218-9 (2022).

479. Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K. & Gilad, Y. A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. *PLoS Genetics* **7,** e1001316. doi:10.1371/journal.pgen.1001316 (2011).

480. Yi, P., Xu, X., Yao, J. & Qiu, B. Analysis of mRNA Expression and DNA Methylation Datasets According to the Genomic Distribution of CpG Sites in Osteoarthritis. *Frontiers in Genetics* **12,** 618803. doi:10.3389/fgene.2021.618803 (2021).

481. Jones, P. A. Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. eng. *Nature Reviews. Genetics* **13,** 484–492. doi:10.1038/nrg3230 (2012).

482. Zhou, S., Shen, Y., Zheng, M., Wang, L., Che, R., Hu, W. & Li, P. DNA Methylation of METTL7A Gene Body Regulates Its Transcriptional Level in Thyroid Cancer. eng. *Oncotarget* **8,** 34652–34660. doi:10.18632/oncotarget.16147 (2017).

483. Teissandier, A. & Bourc'his, D. Gene Body DNA Methylation Conspires with H3K36me3 to Preclude Aberrant Transcription. eng. *The EMBO journal* **36,** 1471–1473. doi:10.15252/embj.201796812 (2017).

484. Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D. & Scandura, J. M. DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. en. *PLOS ONE* **6,** e14524. doi:10.1371/journal.pone.0014524 (2011).

485. Aran, D. & Hellman, A. DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. English. *Cell* **154,** 11–13. doi:10.1016/j.cell.2013.06.018 (2013).

486. Angeloni, A. & Bogdanovic, O. Enhancer DNA Methylation: Implications for Gene Regulation. eng. *Essays in Biochemistry* **63,** 707–715. doi:10.1042/EBC20190030 (2019).

487. Clermont, P.-L., Parolia, A., Liu, H. H. & Helgason, C. D. DNA Methylation at Enhancer Regions: Novel Avenues for Epigenetic Biomarker Development. eng. *Frontiers in Bioscience (Landmark Edition)* **21,** 430–446. doi:10.2741/4399 (2016).

488. Kennedy, A. E., Ozbek, U. & Dorak, M. T. What Has GWAS Done for HLA and Disease Associations? eng. *International Journal of Immunogenetics* **44,** 195–211. doi:10.1111/iji.12332 (2017).

489. Xie, T., Rowen, L., Aguado, B., Ahearn, M. E., Madan, A., Qin, S., Campbell, R. D. & Hood, L. Analysis of the Gene-Dense Major Histocompatibility Complex Class III Region and Its Comparison to Mouse. *Genome Research* **13,** 2621–2636. doi:10.1101/gr.1736803 (2003).

490. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting Genome-Wide DNA Methylation Using Methylation Marks, Genomic Position, and DNA Regulatory Elements. *Genome Biology* **16,** 14. doi:10.1186/s13059-015-0581-9 (2015).

491. Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y. & Pritchard, J. K. DNA Methylation Patterns Associate with Genetic and Gene Expression Variation in HapMap Cell Lines. *Genome Biology* **12,** R10. doi:10.1186/gb-2011-12-1-r10 (2011).

492. Eckhardt, F. *et al.* DNA Methylation Profiling of Human Chromosomes 6, 20 and 22. eng. *Nature Genetics* **38,** 1378–1385. doi:10.1038/ng1909 (2006).

493. Park, S.-M., Choi, E.-Y., Bae, M., Choi, J. K. & Kim, Y.-J. A Long-Range Interactive DNA Methylation Marker Panel for the Promoters of HOXA9 and HOXA10 Predicts Survival in Breast Cancer Patients. *Clinical Epigenetics* **9,** 73. doi:10.1186/s13148-017-0373-z (2017).

494. Kim, S., Park, H. J., Cui, X. & Zhi, D. Collective Effects of Long-Range DNA Methylations Predict Gene Expressions and Estimate Phenotypes in Cancer. *Scientific Reports* **10,** 3920. doi:10.1038/s41598-020-60845-2 (2020).

495. Friman, E. T., Flyamer, I. M., Marenduzzo, D., Boyle, S. & Bickmore, W. A. Ultra-Long-Range Interactions between Active Regulatory Elements. en. *Genome Research* **33,** 1269–1283. doi:10.1101/gr.277567.122 (2023).

496. Gatev, E., Gladish, N., Mostafavi, S. & Kobor, M. S. CoMeBack: DNA Methylation Array Data Analysis for Co-Methylated Regions. *Bioinformatics* **36,** 2675–2683. doi:10.1093/bioinformatics/btaa049 (2020).

497. Mordaunt, C. E., Mouat, J. S., Schmidt, R. J. & LaSalle, J. M. Comethyl: A Network-Based Methylome Approach to Investigate the Multivariate Nature of Health and Disease. en. *bioRxiv,* 2021.07.14.452385. doi:10.1101/2021.07.14.452385 (2021).

498. Johansson, A., Enroth, S. & Gyllensten, U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. eng. *PloS One* **8,** e67378. doi:10.1371/journal.pone.0067378 (2013).

499. Battram, T. *et al.* The EWAS Catalog: A Database of Epigenome-Wide Association Studies. eng. *Wellcome Open Research* **7,** 41. doi:10.12688/wellcomeopenres.17598.2 (2022).

500. Levy, J. J., Titus, A. J., Salas, L. A. & Christensen, B. C. PyMethylProcess—Convenient High-Throughput Preprocessing Workflow for DNA Methylation Data. *Bioinformatics* **35,** 5379–5381. doi:10.1093/bioinformatics/btz594 (2019).

501. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. en. *bioRxiv* (2015).

502. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. en. *arXiv* (2015).

503. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. en. *arXiv* (2017).

504. Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019).

505. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. & Cournapeau, D. Scikit-Learn: Machine Learning in Python. en. *MACHINE LEARNING IN PYTHON,* 6 (2011).

506. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: An R Package for Analyzing Data from Illumina's HumanMethylation450 Platform. eng. *Bioinformatics (Oxford, England)* **32,** 286–288. doi:10.1093/bioinformatics/btv560 (2016).

507. Stranneheim, H. & Wedell, A. Exome and Genome Sequencing: A Revolution for the Discovery and Diagnosis of Monogenic Disorders. *Journal of Internal Medicine* **279,** 3–15. doi:10.1111/joim.12399 (2016).

508. *Doctoral Studies as Part of an Innovative Training Network (ITN)* [Accessed 01-12-2023]. 2023. [Online].%20Available%20from:%20%5Curl%7Bhttps://open-research-europe.ec.europa.eu/articles/1-34%7D.

509. Doonan, F., Taylor, L., Branduardi, P. & Morrissey, J. P. Innovative Training Networks: Overview of the Marie Skłodowska-Curie PhD Training Model. *FEMS Microbiology Letters* **365,** fny207. doi:10.1093/femsle/fny207 (2018).

510. *marie-sklodowska-curie-actions.ec.europa.eu* [Accessed 01-12-2023]. 2023. [Online].%20Available%20from:%20%5Curl%7Bhttps://marie-sklodowska-curie-actions.ec.europa.eu/%7D.

511. *cordis.europa.eu* [Accessed 01-12-2023]. 2023. [Online]. %20Available%20from: %20%5Curl%7Bhttps://cordis.europa.eu/programme/id/H2020_MSCA-ITN-2019%7D.

512. *Translational SYStemics: Personalised Medicine at the Interface of Translational Research and Systems Medicine* [Accessed 01-12-2023]. 2023. [Online].%20Available%20from:%20%5Curl%7Bhttps://cordis.europa.eu/project/id/860895%7D.

513. Morris, S. A., Alsaidi, A. T., Verbyla, A., Cruz, A., Macfarlane, C., Bauer, J. & Patel, J. N. Cost Effectiveness of Pharmacogenetic Testing for Drugs with Clinical Pharmacogenetics Implementation Consortium (CPIC) Guidelines: A Systematic Review. *Clinical Pharmacology and Therapeutics* **112,** 1318–1328. doi:10.1002/cpt.2754 (2022).

514. Bienfait, K., Chhibber, A., Marshall, J.-C., Armstrong, M., Cox, C., Shaw, P. M. & Paulding, C. Current challenges and opportunities for pharmacogenomics: perspective of the Industry Pharmacogenomics Working Group (I-PWG). *Human Genetics* **141,** 1165–1173 (2022).

515. Najjary, S., Kros, J. M., Stricker, B. H., Ruiter, R., Shuai, Y., Kraaij, R., Van Steen, K., van der Spek, P., Van Eijck, C. H., Ikram, M. A., *et al.* Association of blood cell-based inflammatory markers with gut microbiota and cancer incidence in the Rotterdam study. *Cancer Medicine* (2023).

516. Walakira, A., Rozman, D., Režen, T., Mraz, M. & Moškon, M. Guided Extraction of Genome-Scale Metabolic Models for the Integration and Analysis of Omics Data. *Computational and Structural Biotechnology Journal* **19,** 3521–3530. doi:10.1016/j.csbj.2021.06.009 (2021).

517. Walakira, A., Skubic, C., Nadižar, N., Rozman, D., Režen, T., Mraz, M. & Moškon, M. Integrative Computational Modeling to Unravel Novel Potential Biomarkers in Hepatocellular Carcinoma. *Computers in Biology and Medicine* **159,** 106957. doi:10.1016/j.compbiomed.2023.106957 (2023).

518. Melograna, F., Li, Z., Galazzo, G., van Best, N., Mommers, M., Penders, J., Stella, F. & Van Steen, K. Edge and Modular Significance Assessment in Individual-Specific Networks. en. *Scientific Reports* **13,** 7868. doi:10.1038/s41598-023-34759-8 (2023).

519. Van Hilten, A., Melograna, F., Fan, B., Niessen, W. J., van Steen, K. & Roshchupkin, G. V. Detecting Genetic Interactions with Visible Neural Networks. *bioRxiv,* 2024–02. doi:doi.org/10.1101/2024.02.27.582086 (2024).

520. Li, Z., Melograna, F., Hoskens, H., Duroux, D., Marazita, M. L., Walsh, S., Weinberg, S. M., Shriver, M. D., Müller-Myhsok, B., Claes, P. & Van Steen, K. netMUG: A Novel Network-Guided Multi-View Clustering Workflow for Dissecting Genetic and Facial Heterogeneity. *Frontiers in Genetics* **14** (2023).

521. Andreoli, L., Peeters, H., Van Steen, K. & Dierickx, K. Taking the risk. A systematic review of ethical reasons and moral arguments in the clinical use of polygenic risk scores. *American Journal of Medical Genetics Part A,* e63584. doi:doi.org/10.1002/ajmg.a.63584 (2024).

522. Yousefi, B., Melograna, F., Galazzo, G., van Best, N., Mommers, M., Penders, J., Schwikowski, B. & Van Steen, K. Capturing the Dynamics of Microbial Interactions through Individual-Specific Networks. *Frontiers in Microbiology* **14** (2023).

523. Yousefi, B., Firoozbakht, F., Melograna, F., Schwikowski, B. & Van Steen, K. PLEX.I: A Tool to Discover Features in Multiplex Networks That Reflect Clinical Variation. *Frontiers in Genetics* **14** (2023).

524. Yousefi, B. & Schwikowski, B. Consensus Clustering for Robust Bioinformatics Analysis. *bioRxiv,* 2024–03. doi:doi.org/10.1101/2024.03.21.586064 (2024).

525. Mihajlović, K., Ceddia, G., Malod-Dognin, N., Novak, G., Kyriakis, D., Skupin, A. & Pržulj, N. Multi-Omics Integration of scRNA-Seq Time Series Data Predicts New Intervention Points for Parkinson's Disease. *bioRxiv,* 2023–12 (2023).

526. Gureghian, V. *et al.* A Multi-Omics Integrative Approach Unravels Novel Genes and Pathways Associated with Senescence Escape after Targeted Therapy in NRAS Mutant Melanoma. en. *Cancer Gene Therapy* **30,** 1330–1345. doi:10.1038/s41417-023-00640-z (2023).

527. Li, S., Schmid, K. T., de Vries, D. H., Korshevniuk, M., Losert, C., Oelen, R., van Blokland, I. V., Groot, H. E., Swertz, M. A., van der Harst, P., Westra, H.-J., van der Wijst, M. G., Heinig, M., Franke, L. & BIOS Consortium, e. C. Identification of Genetic Variants That Impact Gene Co-Expression Relationships Using Large-Scale Single-Cell Data. en. *Genome Biology* **24,** 80. doi:10.1186/s13059-023-02897-x (2023).

528. Knauer-Arloth, J., Hryhorzhevska, A. & Binder, E. B. Multi-Omics Analysis of the Molecular Response to Glucocorticoids-Insights into Shared Genetic Risk from Psychiatric to Medical Disorders. *medRxiv,* 2023–12 (2023).

529. Skokou, M., Karamperis, K., Koufaki, M.-I., Tsermpini, E.-E., Pandi, M.-T., Siamoglou, S., Ferentinos, P., Bartsakoulia, M., Katsila, T., Mitropoulou, C., *et al.* Clinical Implementation of Preemptive Pharmacogenomics in Psychiatry. *Ebiomedicine* **101** (2024).

530. Karamperis, K., Koromina, M., Papantoniou, P., Skokou, M., Kanellakis, F., Mitropoulos, K., Vozikis, A., Müller, D. J., Patrinos, G. P. & Mitropoulou, C. Economic Evaluation in Psychiatric Pharmacogenomics: A Systematic Review. en. *The Pharmacogenomics Journal* **21,** 533–541. doi:10.1038/s41397-021-00249-1 (2021).

531. Swen, J. J., van der Wouden, C. H., Manson, L. E., Abdullah-Koolmees, H., Blagec, K., Blagus, T., Böhringer, S., Cambon-Thomsen, A., Cecchin, E., Cheung, K.-C., *et al.* A 12-Gene Pharmacogenetic Panel to Prevent Adverse Drug Reactions: An Open-Label, Multicentre, Controlled, Cluster-Randomised Crossover Implementation Study. *The Lancet* **401,** 347–356 (2023).

532. Katz, S., Martins dos Santos, V. A., Saccenti, E. & Roshchupkin, G. V. mEthAE: An Explainable AutoEncoder for Methylation Data. *bioRxiv,* 2023–07 (2023).

533. Li, Z., Katz, S., Martins dos Santos, V. A., Fardo, D., Claes, P., Saccenti, E., Van Steen, K. & Roshchupkin, G. V. Novel Multi-Omics Deconfounding Variational Autoencoders Can Obtain Meaningful Disease Subtyping. *bioRxiv,* 2024–02 (2024).

534. Kočar, E., Katz, S., Pušnik, Ž., Bogovič, P., Turel, G., Skubic, C., Režen, T., Strle, F., Dos Santos, V. A. M., Mraz, M., *et al.* COVID-19 and Cholesterol Biosynthesis: Towards Innovative Decision Support Systems. *Iscience* **26** (2023).

535. Stenzinger, A. *et al.* Trailblazing Precision Medicine in Europe: A Joint View by Genomic Medicine Sweden and the Centers for Personalized Medicine, ZPM, in Germany. *Seminars in Cancer Biology* **84,** 242–254. doi:10.1016/j.semcancer.2021.05.026 (2022).

536. Lévy, Y. Genomic Medicine 2025: France in the Race for Precision Medicine. *The Lancet* **388,** 2872. doi:10.1016/S0140-6736(16)32467-9 (2016).

537. Stenzinger, A. *et al.* Implementation of Precision Medicine in Healthcare—A European Perspective. *Journal of Internal Medicine* **294,** 437–454. doi:10.1111/joim.13698 (2023).

538. Bedard, P. L., Hyman, D. M., Davids, M. S. & Siu, L. L. Small Molecules, Big Impact: 20 Years of Targeted Therapy in Oncology. *Lancet (London, England)* **395,** 1078–1088. doi:10.1016/S0140-6736(20)30164-1 (2020).

539. Rosenquist, R., Fröhling, S. & Stamatopoulos, K. Precision Medicine in Cancer: A Paradigm Shift. *Seminars in Cancer Biology* **84,** 1–2. doi:10.1016/j.semcancer.2022.05.008 (2022).

540. Commission), D.-G. f. R.
bibinitperiod I. (*et al. Evaluation Study of the European Framework Programmes for Research and Innovation for Excellent Science: Horizon 2020 : Phase 1 Final Study Report* ISBN: 978-92-68-01934-4 (Publications Office of the European Union, LU, 2023).

541. *Yeast Cell Factories: Training Researchers to Apply Modern Post-Genomic Methods In Yeast Biotechnology — cordis.europa.eu.* [Accessed 01-12-2023]. 2023. [Online].%20Available%20from:%20%5Curl%7Bhttps://cordis.europa.eu/project/id/606795%7D.

542. Woolston, C. Depression and Anxiety 'the Norm' for UK PhD Students. *Nature.* doi:10.1038/d41586-021-03761-3 (2021).

543. Visvikis-Siest, S., Theodoridou, D., Kontoe, M.-S., Kumar, S. & Marschler, M. Milestones in Personalized Medicine: From the Ancient Time to Nowadays—the Provocation of COVID-19. *Frontiers in Genetics* **11,** 569175. doi:10.3389/fgene.2020.569175 (2020).

544. Faulkner, E. *et al.* Being Precise about Precision Medicine: What Should Value Frameworks Incorporate to Address Precision Medicine? A Report of the Personalized Precision Medicine Special Interest Group. *Value in Health* **23,** 529–539. doi:10.1016/j.jval.2019.11.010 (2020).

545. Suijker, J., Pijpe, A., Hoogerbrug, D., Heymans, M. W., van Zuijlen, P. P. M., Halm, J. A., Group, N. K. C. & Meij-de Vries, A. IDENTIFICATION OF POTENTIALLY MODIFIABLE FACTORS TO IMPROVE RECOGNITION AND OUTCOME OF NECROTIZING SOFT-TISSUE INFECTIONS. en-US. *Shock* **61,** 585. doi:10.1097/SHK.0000000000002325 (2024).

546. Fernandes, M., Vieira, S. M., Leite, F., Palos, C., Finkelstein, S. & Sousa, J. M. C. Clinical Decision Support Systems for Triage in the Emergency Department Using Intelligent Systems: A Review. eng. *Artificial Intelligence in Medicine* **102,** 101762. doi:10.1016/j.artmed.2019.101762 (2020).

547. Du, Y., McNestry, C., Wei, L., Antoniadi, A. M., McAuliffe, F. M. & Mooney, C. Machine Learning-Based Clinical Decision Support Systems for Pregnancy Care: A Systematic Review. eng. *International Journal of Medical Informatics* **173,** 105040. doi:10.1016/j.ijmedinf.2023.105040 (2023).

548. Theodosiou, A. A. & Read, R. C. Artificial Intelligence, Machine Learning and Deep Learning: Potential Resources for the Infection Clinician. eng. *The Journal of Infection* **87,** 287–294. doi:10.1016/j.jinf.2023.07.006 (2023).

549. Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., Buchard, A., Georgiou, P., Lescure, F.-X., Birgand, G. & Holmes, A. H. Machine Learning for Clinical Decision Support in Infectious Diseases: A Narrative Review of Current Applications. eng. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **26,** 584–595. doi:10.1016/j.cmi.2019.09.009 (2020).

550. Schukat, M., McCaldin, D., Wang, K., Schreier, G., Lovell, N. H., Marschollek, M. & Redmond, S. J. Unintended Consequences of Wearable Sensor Use in Healthcare. *Yearbook of Medical Informatics,* 73–86. doi:10.15265/IY-2016-025 (2016).

551. statista. *Direct-to-Consumer Genetic Testing Market Size Worldwide 2014-2022 | Statista* 2024. https://www.statista.com/statistics/792022/global-direct-to-consumer-genetic-testing-market-size/.

552. Intelligence, G. T. *Data-Driven Approach to Fitness Has Become Obsessional* en-US. 2022.

553. Ruhl, G. L., Hazel, J. W., Clayton, E. W. & Malin, B. A. Public Attitudes Toward Direct to Consumer Genetic Testing. *AMIA Annual Symposium Proceedings* **2019,** 774–783 (2020).

554. Hedetoft, M., Madsen, M. B., Madsen, L. B. & Hyldegaard, O. Incidence, Comorbidity and Mortality in Patients with Necrotising Soft-Tissue Infections, 2005–2018: A Danish Nationwide Register-Based Cohort Study. *BMJ Open* **10,** e041302. doi:10.1136/bmjopen-2020-041302 (2020).

555. Suijker, J., Troncoso, E., Pizarro, F., Montecinos, S., Villarroel, G., Erazo, C., Cisternas, J. P., Andrades, P., Benítez, S., Sepúlveda, S. & Danilla, S. Long-Term Quality-of-Life Outcomes after Body Contouring Surgery: Phase IV Results for the Body-QoL® Cohort. *Aesthetic Surgery Journal* **38,** 279–288. doi:10.1093/asj/sjx090 (2017).

556. Suijker, J., Stoop, M., Meij-de Vries, A., Pijpe, A., Boekelaar, A., Egberts, M. & Van Loey, N. The Impact of Necrotizing Soft Tissue Infections on the Lives of Survivors: A Qualitative Study. en. *Quality of Life Research* **32,** 2013–2024. doi:10.1007/s11136-023-03371-8 (2023).

557. Grimshaw, J. M., Patey, A. M., Kirkham, K. R., Hall, A., Dowling, S. K., Rodondi, N., Ellen, M., Kool, T., van Dulmen, S. A., Kerr, E. A., Linklater, S., Levinson, W. & Bhatia, R. S. De-Implementing Wisely: Developing the Evidence Base to Reduce Low-Value Care. *BMJ Quality & Safety* **29,** 409–417. doi:10.1136/bmjqs-2019-010060 (2020).

558. Mohsin, S. N., Gapizov, A., Ekhator, C., Ain, N. U., Ahmad, S., Khan, M., Barker, C., Hussain, M., Malineni, J., Ramadhan, A. & Halappa Nagaraj, R. The Role of Artificial Intelligence in Prediction, Risk Stratification, and Personalized Treatment Planning for Congenital Heart Diseases. eng. *Cureus* **15,** e44374. doi:10.7759/cureus.44374 (2023).

559. Athieniti, E. & Spyrou, G. M. A Guide to Multi-Omics Data Collection and Integration for Translational Medicine. *Computational and Structural Biotechnology Journal* **21,** 134–149. doi:10.1016/j.csbj.2022.11.050 (2022).

560. Beaulieu-Jones, B. K., Yuan, W., Brat, G. A., Beam, A. L., Weber, G., Ruffin, M. & Kohane, I. S. Machine Learning for Patient Risk Stratification: Standing on, or Looking over, the Shoulders of Clinicians? en. *npj Digital Medicine* **4,** 1–6. doi:10.1038/s41746-021-00426-3 (2021).

561. Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. en. *npj Digital Medicine* **3,** 1–10. doi:10.1038/s41746-020-0221-y (2020).

562. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing Bias in Big Data and AI for Health Care: A Call for Open Science. *Patterns* **2,** 100347. doi:10.1016/j.patter.2021.100347 (2021).

563. Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres Terre, H., Shams, Z., Jamnik, M. & Liò, P. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *Frontiers in Genetics* **10** (2019).

564. Lawry Aguila, A., Chapman, J., Janahi, M. & Altmann, A. en. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (eds Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) 430–440 (Springer Nature Switzerland, Cham, 2022). ISBN: 978-3-031-16430-9 978-3-031-16431-6. doi:10.1007/978-3-031-16431-6_41.

565. Liu, X., Li, B., Bron, E., Niessen, W., Wolvius, E. & Roshchupkin, G. Projection-Wise Disentangling for Fair and Interpretable Representation Learning: Application to 3D Facial Shape Analysis. *arXiv:2106.13734 [cs]* (2021).

566. Morgenpost. *Corona-Lage in Kliniken: Intensivbetten Und Fälle Auf Normalstationen* 2024. https://interaktiv.morgenpost.de/corona-deutschland-intensiv-betten-monitor-krankenhaus-auslastung/.

567. Marckmann, G., Neitzke, G. & Schildmann, J. Triage in Der COVID-19-Pandemie–Was Ist Gerecht. *DIVI* **11,** 172–178 (2020).

568. Ventola, C. L. The Antibiotic Resistance Crisis. *Pharmacy and Therapeutics* **40,** 277–283 (2015).

569. DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. The Price of Innovation: New Estimates of Drug Development Costs. *Journal of Health Economics* **22,** 151–185. doi:10.1016/S0167-6296(02)00126-1 (2003).

570. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the Decline in Pharmaceutical R&D Efficiency. en. *Nature Reviews Drug Discovery* **11,** 191–200. doi:10.1038/nrd3681 (2012).

571. Ginesta Roque, L. *Neurologische Nebenwirkungen einer Immuncheckpoint-Therapie* ger. PhD thesis (Charité - Universitätsmedizin Berlin, 2023).

572. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the Black Box: Interpretable Machine Learning for Geneticists. English. *Trends in Genetics* **36,** 442–455. doi:10.1016/j.tig.2020.03.005 (2020).

573. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A Systematic Review of Biologically-Informed Deep Learning Models for Cancer: Fundamental Trends for Encoding and Interpreting Oncology Data. *BMC Bioinformatics* **24,** 198. doi:10.1186/s12859-023-05262-8 (2023).

574. Cesario, A., Auffray, C., Russo, P. & Hood, L. P4 Medicine Needs P4 Education. en. *Current Pharmaceutical Design* **20,** 6071–6072. doi:10.2174/138161282066140314145445.

575. explodingtopics. *Number of ChatGPT Users (May 2024)* en. 2023. https://explodingtopics.com/blog/chatgpt-users.

576. Hu, K. & Hu, K. ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note. en. *Reuters* (2023).

577. Wikipedia. Epic Systems. en. *Wikipedia* (2024).

578. Creswell, J. Doctors Find Barriers to Sharing Digital Medical Records. en-US. *The New York Times* (2014).

579. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for Scientific Data Management and Stewardship. en. *Scientific Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).

580. Turki, H., Rasberry, L., Hadj Taieb, M. A., Mietchen, D., Ben Aouicha, M., Pouris, A. & Bousrih, Y. Letter to the Editor: FHIR RDF - Why the World Needs Structured Electronic Health Records. *Journal of Biomedical Informatics* **136,** 104253. doi:10.1016/j.jbi.2022.104253 (2022).

581. Touré, V., Krauss, P., Gnodtke, K., Buchhorn, J., Unni, D., Horki, P., Raisaro, J. L., Kalt, K., Teixeira, D., Crameri, K. & Österle, S. FAIRification of Health-Related Data Using Semantic Web Technologies in the Swiss Personalized Health Network. en. *Scientific Data* **10,** 127. doi:10.1038/s41597-023-02028-y (2023).

582. Lin, A. Y. *et al.* Improving the Quality and Utility of Electronic Health Record Data through Ontologies. en. *Standards* **3,** 316–340. doi:10.3390/standards3030023 (2023).

583. Nijsse, B., Schaap, P. J. & Koehorst, J. J. FAIR Data Station for Lightweight Metadata Management and Validation of Omics Studies. *GigaScience* **12,** giad014. doi:10.1093/gigascience/giad014 (2023).

584. Datta, B. N. *Numerical Linear Algebra and Applications* (SIAM, 2010).

585. *An Empirical Analysis of Compute-Optimal Large Language Model Training* en. 2022. https://deepmind.google/discover/blog/an-empirical-analysis-of-compute-optimal-large-language-model-training/.

586. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N. & Wang, B. scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI. en. *Nature Methods,* 1–11. doi:10.1038/s41592-024-02201-0 (2024).

587. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature machine intelligence* **1,** 206–215 (2019).

588. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. & Zhong, C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys* **16,** 1–85. doi:10.1214/21-SS133 (2022).

589. Rebecca Boiarsky, Nalini Singh, Alejandro Buendia, Gad Getz & David Sontag. A Deep Dive into Single-Cell RNA Sequencing Foundation Models. *bioRxiv,* 2023.10.19.563100. doi:10.1101/2023.10.19.563100 (2023).

590. Gosiewska, A., Kozak, A. & Biecek, P. Simpler Is Better: Lifting Interpretability-Performance Trade-off via Automated Feature Engineering. *Decision Support Systems. Interpretable Data Science For Decision Making* **150,** 113556. doi:10.1016/j.dss.2021.113556 (2021).

591. Liao, L., Li, H., Shang, W. & Ma, L. An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. *ACM Transactions on Software Engineering and Methodology* **31,** 53:1–53:40. doi:10.1145/3506695 (2022).

592. Johansson, U., Sönströd, C., Norinder, U. & Boström, H. Trade-off between Accuracy and Interpretability for Predictive in Silico Modeling. eng. *Future Medicinal Chemistry* **3,** 647–663. doi:10.4155/fmc.11.23 (2011).

593. Causality in Digital Medicine. en. *Nature Communications* **12,** 5471. doi:10.1038/s41467-021-25743-9 (2021).

594. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome Signature Genes in Breast Cancer: Is There a Unique Set? *Bioinformatics* **21,** 171–178. doi:10.1093/bioinformatics/bth469 (2005).

595. Fröhlich, H. *et al.* From Hype to Reality: Data Science Enabling Personalized Medicine. *BMC Medicine* **16,** 150. doi:10.1186/s12916-018-1122-7 (2018).

596. Heinze-Deml, C., Maathuis, M. H. & Meinshausen, N. Causal Structure Learning. en. *Annual Review of Statistics and Its Application* **5,** 371–391. doi:10.1146/annurev-statistics-031017-100630 (2018).

597. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q. & Tsaftaris, S. A. Causal Machine Learning for Healthcare and Precision Medicine. *Royal Society Open Science* **9,** 220638. doi:10.1098/rsos.220638 (2022).

598. Choo, D., Shiragur, K. & Uhler, C. Causal Discovery under Off-Target Interventions. *arXiv* (2024).

599. Rathnam, C., Lee, S. & Jiang, X. An Algorithm for Direct Causal Learning of Influences on Patient Outcomes. *Artificial Intelligence in Medicine* **75,** 1–15. doi:10.1016/j.artmed.2016.10.003 (2017).

600. Baker, R. E., Peña, J.-M., Jayamohan, J. & Jérusalem, A. Mechanistic Models versus Machine Learning, a Fight Worth Fighting for the Biological Community? *Biology Letters* **14,** 20170660. doi:10.1098/rsbl.2017.0660 (2018).

601. Grant, J., Hua, F., Apgar, J. F., Burke, J. M. & Marcantonio, D. H. Mechanistic PK/PD Modeling to Address Early-Stage Biotherapeutic Dosing Feasibility Questions. *mAbs* **15,** 2192251. doi:10.1080/19420862.2023.2192251 (2023).

602. Gotsmy, M., Brunmair, J., Büschl, C., Gerner, C. & Zanghellini, J. Probabilistic Quotient's Work and Pharmacokinetics' Contribution: Countering Size Effect in Metabolic Time Series Measurements. en. *BMC Bioinformatics* **23,** 379. doi:10.1186/s12859-022-04918-1 (2022).

603. Passi, A., Tibocha-Bonilla, J. D., Kumar, M., Tec-Campos, D., Zengler, K. & Zuniga, C. Genome-Scale Metabolic Modeling Enables In-Depth Understanding of Big Data. *Metabolites* **12,** 14. doi:10.3390/metabo12010014 (2021).

604. Libiseller-Egger, J., Coltman, B. L., Gerstl, M. P. & Zanghellini, J. Environmental Flexibility Does Not Explain Metabolic Robustness. en. *npj Systems Biology and Applications* **6,** 1–9. doi:10.1038/s41540-020-00155-5 (2020).

605. Bentzen, H. B. Exchange of Human Data Across International Boundaries. en. *Annual Review of Biomedical Data Science* **5,** 233–250. doi:10.1146/annurev-biodatasci-122220-110811 (2022).

606. edpb. *Recommendations 01/2020 on Measures That Supplement Transfer Tools to Ensure Compliance with the EU Level of Protection of Personal Data | European Data Protection Board* https://www.edpb.europa.eu/our-work-tools/our-documents/recommendations/recommendations-012020-measures-supplement-transfer%5C_en.

607. Shiranthika, C., Saeedi, P. & Bajić, I. V. Decentralized Learning in Healthcare: A Review of Emerging Techniques. *IEEE Access* **11,** 54188–54209. doi:10.1109/ACCESS.2023.3281832 (2023).

608. Joshi, M., Pal, A. & Sankarasubbu, M. Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. *ACM Trans. Comput. Healthcare* **3.** doi:10.1145/3533708 (2022).

609. Dhade, P. & Shirke, P. Federated Learning for Healthcare: A Comprehensive Review. en. *Engineering Proceedings* **59,** 230. doi:10.3390/engproc2023059230 (2024).

610. Berlin, C.-U. *Strategic Direction 2030* en. 2024. https://www.charite.de/en/charite/about%5C_us/strategic%5C_direction%5C_2030/.

611. Ärzteblatt Redaktion Deutsches, D. Ä. G. *Digitalisierung im Krankenhaus: Der Infrastruktur fehlt die Finanzierung* de. 2017. https://www.aerzteblatt.de/archiv/195006/Digitalisierung-im-Krankenhaus-Der-Infrastruktur-fehlt-die-Finanzierung.

612. Prensky, M. H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom. en. *Innovate: Journal of Online Education* **5** (2009).

List of Publications

This thesis:

Katz S, Suijker J, Hardt C, Madsen MB, Meij-de Vries A, Pijpe A, Skrede S, Hyldegaard O, Solligard E, Norrby-Teglund A, Saccenti E. Decision support system and outcome prediction in a cohort of patients with necrotizing soft-tissue infections. *International journal of medical informatics. 2022 Nov 1;167:104878.*

Kočar E, **Katz S**, Pušnik Ž, Bogovič P, Turel G, Skubic C, Režen T, Strle F, Dos Santos VA, Mraz M, Moškon M. COVID-19 and cholesterol biosynthesis: Towards innovative decision support systems. *Iscience. 2023 Oct 20;26(10)*.

Katz S, Suijker J, Skrede S, Meij-de Vries A, Pijpe A, Norrby-Teglund A, Palma Medina LM, Damas JK, Hyldegaard O, Solligard E, Svensson M, Anders Mosevoll K, Martins dos Santos VA, Saccenti E. A validated model for early prediction of group A streptococcal aetiology and clinical endpoints in necrotising soft tissue infections. *Manuscript submitted.*

Li Z, **Katz S**, Saccenti E, Fardo DW, Claes P, Martins dos Santos VA, Van Steen K, Roshchupkin GV. Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping. *bioRxiv. 2024 Feb 8:2024-02.*

Van Hilten A, **Katz S**, Saccenti E, Niessen WJ, Roshchupkin GV. Designing Interpretable Deep Learning Applications for Functional Genomics: a Quantitative Analysis. *Manuscript submitted.*

Katz S, Martins dos Santos VA, Saccenti E, Roshchupkin GV. mEthAE: an Explainable AutoEncoder for methylation data. *bioRxiv. 2023 Jul 19:2023-07.*

Katz S, Andreoli L, Berca C, Korshevniuk M, Head RM, Van Steen K. Bridging the Gap in Precision Medicine: TranSYS Training Programme for Next-Generation Scientists. *Frontiers in Medicine.;11:1348148*.

Overview of the Completed Training Activities

9.4 Discipline specific activities

Name of the course/meeting	Organizing institute(s)	City (Country)	Year
TranSYS Summer School 1	UL / TranSYS	Ljubljana (SLO)	2020
TranSYS Summer School 2	Insititute Pasteur / TranSYS	Paris (FRA)	2021
TranSYS Summer School 3	University of Patras / TranSYS	Patras (GRC)	2022
TranSYS Bootcamp 1	EMC / TranSYS	Rotterdam (NL)	2022
TranSYS Bootcamp 2	KUL / TranSYS	Leuven (BE)	2023
Integrated Modeling and Optimization	BioSB	online	2020
BioSB 2021	BioSB	online	2021
BioSB 2022	BioSB	Lunteren (NL)	2022
Future Medicine Science Match 2020	BIH	Berlin (GER)	2020
BioHackathon ELIXIR	EXLIXIR	Barcelona (ESP)	2021
RECOMB 2022	RECOMB	San Francisco (USA)	2022
Epigenomics of Common Disease	Wellcome Connecting Science	Cambridge (UK)	2022
EMBL 2024	EMBL	Heidelberg (GER)	2024
ESHG 2024	ESHG	Berlin (GER)	2024

9.5 General courses

Name of the course/meeting	Organizing institute(s)	City (Country)	Year
VLAG PhD week	VLAG	Baarlo (NL)	2021
Efficient writing strategies	WGS	online	2021
Scientific Writing	WGS	online	2021
Project and Time Management	WGS	online	2021
Scientific Artwork, Data visualisation and Infographics with Adobe Illustrator	WGS	online	2022

9.6 Other activities

Name of the course/meeting	Organizing institute(s)	City (Country)	Year
Preparation of research proposal	SSB	Wageningen (NL)	2020
PhD trip	SSB	USA	2022
TranSYS Journal Club	TranSYS	online	2021-2022
SSB Journal Club & Seminar	SSB	Wageningen (NL)	2020-2024

About the Author

Sonja Katz was born on June 16, 1995 in St. Veit an der Glan, Austria. She pursued a BSc in Food Science and Biotechnology and a MSc in Biotechnology at the University of Natural Resources and Life Sciences in Vienna, Austria. In 2020, Sonja became an early stage researcher (ESR) under the frame of the Innovative Training Network "*Translational SYStemics: Personalised Medicine at the Interface of Translational Research and Systems Medicine (TranSYS)*" under the beneficiary Lifeglimmer GmbH, Berlin, Germany, and affiliated as external PhD candidate with the System and Synthetic Biology group at Wageningen University. From 2023 until the end of her PhD she was employed at the Erasmus University Medical Center Rotterdam, The Netherlands, as a research scientist.
Her research focused on the development of artificial intelligence models and methodologies to capture patient characteristics utilising a variety of data, with the ultimate goal of enhancing personalised treatment approaches and patient outcomes. In an attempt to bridge the gap between research and clinical practice she prioritised the explainability of models as well as disseminate the results of her research in the form of clinical decision support systems, which are built to be easily integratable in daily clinical practices. The findings of her research are presented in this thesis.

Acknowledgements

Despite all my ambitious journalling and note-keeping efforts over the past years to track my research, life goals, and personal development, which have all failed, the list of people I am grateful for in enabling this milestone in my life was the only thing I managed to continuously fill throughout these years. By the time you are reading this, however, I am sure I will be kicking myself because I will have forgotten someone - potentially you. So, if you feel like your name should be here and you can't find it, don't worry; you were not forgotten, and I am already busy kicking myself.

Firstly, I want to thank the people who set me on my academic path and showed me that computational biology can be more than just aligning genomes to references (no offense intended). Especially **Chris Oostenbrink**, who, for me, still very much embodies the academic spirit as it should be. He gave me - a random student approaching him - a computer, a dauntingly thick manual, and the chance to explore something new. Look at the mess you caused, Chris, and thank you for it. Of course, without the collaborators at MMS, I would have i) never managed the project at the time and ii) probably died of imposter syndrome. So special thanks go to **Matthias** and **Öhli**, for supporting me and monitoring my development as closely as my cluster jobs, which miraculously (!) always seemed to crash. Reminiscent of the past, I would also like to take the opportunity to thank **Peter Sykacek** and **Thomas Mohr**, who have demonstrated that offering encouraging words (or an email) and providing an opportunity to learn doesn't cost much but can significantly impact someone's path in life.
I would have never come this far in my academic career (being serious here) if I hadn't met the amazing group of people I studied with. The initial carbon group merged with chlorine and their nuclear fusion resulted in the Chemistry Catz, the most stable element discovered thus far. Dear **Marlene**, **Michi**, **Johnny**, **Mathias**, and **Julian** - despite our paths diverging, I am pleasantly surprised by our constant efforts to stay in touch and help each other. I hope we keep this up for many years to come. Special gratitude to **Julian**, my personal

Kummerkasten, Oxford-English-Grammar-Checker (now mostly replaced by ChatGPT due to lower labour costs), and close companion for nearly a decade now. Without you, I would have never had the courage (or the grades) to embark on this journey.
I want to extend my gratitude to the people who guided me, especially at the beginning of my PhD. Thank you, **Vitor**, for giving me the opportunity to pursue a PhD despite not having a perfect CV for the position. It was certainly not without drama, but it surely contributed to my character building, for which I am thankful. Special thanks also go to my colleagues from LifeGlimmer: **Lorna**, **Erno**, **Chris**, **Mechthild**, and **Marco**, who welcomed me with open arms and quickly made me feel like a valued team member - one couldn't have asked for a smoother transition into their first job! Especially **Marco**, whom I am honoured to call my paranymph and with whom I shared all the ups and downs of this journey. Without you there listening to my problems and fears, I would have probably quit this PhD halfway through. Thank you, Marco, for your endless positivity and caring personality. Please never change.
Special gratitude also goes to **Jaco**, who bravely accepted the challenge of sharing a co-authorship with me, a greenhorn PhD candidate. Having you, an experienced researcher, by my side was a great relief, and I always felt acknowledged, valued, and taken seriously. Our collaboration was not only a positive professional experience but also led to the blossoming of a friendship; win-win I would say. You also broadened my perspective by introducing me to the clinical aspects of our work and teaching me a new language, which I believe will be invaluable in my future career. Thank you for taking me under your wing and making me your medical apprentice, Jaco!
Among those essential for the successful completion of this PhD, my fellow **TranSYS ESRs** hold a special place. I believe I speak for all of us when I say that none of us fully anticipated what we were getting into with this ITN. However, for us, the cliché saying "it's about the friends we make along the way" truly applies. Despite the challenges posed by the pandemic, we managed to form a supportive peer group. This group not only provided emotional support but we actually managed to get some research done as well. In this spirit, I extend special thanks to my direct collaborators, including **Zuqi**, who patiently endured my poorly debugged Python code, and to **Lara**, **Catalina**, and **Maryna** for our collaborative efforts in writing a paper we are proud of, despite the obstacles we faced. I also want to express my heartfelt gratitude to **Giada** and **Katarina** for the memorable parties in Leuven, indulgent culinary experiences in Barcelona, and for being strong, independent, and supportive females. Every woman would be fortunate to have you by their side.
I also want to thank the SSB group at WUR, especially **Sara M.**, **Sara B.**, **María**, **Enrique**, **Lyon**, **Marco**, **Sanjee**, **Sabine**, **Willemijn**, and **Edoardo**. Thank you for making me feel

like a part of WUR despite living 600 km away, and for the wonderful memories we share from our PhD trip! Special thanks to **Edoardo** for commenting on manuscripts quicker than I can write them and for his natural intuition in when to provide active support, which often results in swiftly solving problems or preventing them altogether.
I want to express my gratitude to Computational Population Biology at EMC, including **Gennady**, **Tareq**, **Arno**, **Jing**, **Lau**, **Rafael**, **Sara**, and **Kiefer**, for welcoming me into their group. I couldn't have been luckier with your habit of adopting stray PhD students for collaborations and ultimately keeping them. This experience has broadened my methodological horizons, shaped my research interests, and provided invaluable emotional support through countless stand-up coffee meetings (which we never managed to keep to the scheduled 20 minutes, not even once) and group meetings. Special gratitude goes to **Gennady**, who, besides his seemingly endless ideas for new and interesting research, has earned his place in my V.I.M. (Very Important Mentor) list through his remarkable ability to see the potential in others and believe in them, even when they themselves rarely do (but they're working on it, I promise).
Last but absolutely not least, I want to thank my family and friends, who all share the common trait of not fully understanding what the hell I've actually been doing these past four years (despite our annual conversation at Christmas dinner, which always begins with the question "Okay, Sonja, also *was* arbeitest du jetzt genau...?"). Nevertheless, they have been incredibly supportive throughout the entire process. Liebe **Familie**, danke für euer offenes Ohr, eure emotionale Unterstützung und die Gewissenheit, immer Heim kommen zu können, egal was passiert.
Special thanks also to the friends I was fortunate to make in Berlin including **Lukas**, **Sasha**, **Lorena**, **Bosse**, the present and former members of Stahli 4., **Marius** and **Santiago**, the Wedding folks, and my volleyball mates.
Without you, **Lukas**, I would have never completed this PhD. Thank you for taking me in, showing me Berlin, introducing me to your friends - who eventually became mine - and being my best friend throughout this tumultuous journey of professional and personal growth. Like my family, you create a sense of home where I can always share my feelings, and for that, I am forever grateful. However, I fear I don't express this often enough, so please know that I value you immensely.
Sasha, you might find it hard to believe, but I'm actually a very critical person, not easily convinced and even more difficult to motivate. However, you have managed to broaden my horizons with new perspectives, adventures, and activities. Your genuineness constantly spark my curiosity, motivating me to seek new experiences, slow down in life, and to not worry so much all the time. Thank you for being the inspiring, curious, and captivating

person you are, and for choosing to walk by my side.
I also want to express my gratitude to **Lorena** for always radiating energy and positivity. You've been my personal motivator for bike tours and ice cream and ready to be ultralight anytime required. You've made my darkest days feel a little brighter every time.

The research described in this thesis was financially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No. 860895 (TranSYS).

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Printed by Proefschriften.nl