

MEETING REPORT

Open Access



# Computational Applications in Secondary Metabolite Discovery (CAiSMD): an online workshop

Fidele Ntie-Kang<sup>1,2,3\*</sup>, Kiran K. Telukunta<sup>4</sup>, Serge A. T. Fobofou<sup>5</sup>, Victor Chukwudi Osamor<sup>6</sup>, Samuel A. Egieyeh<sup>7,8</sup>, Marilia Valli<sup>9</sup>, Yannick Djoumbou-Feunang<sup>10</sup>, Maria Sorokina<sup>11</sup>, Conrad Stork<sup>12</sup>, Neann Mathai<sup>13</sup>, Paul Zierep<sup>14</sup>, Ana L. Chávez-Hernández<sup>15</sup>, Miquel Duran-Frigola<sup>16,17</sup>, Smith B. Babiaka<sup>1</sup>, Romuald Tematio Fouedjou<sup>18</sup>, Donatus B. Eni<sup>1</sup>, Simeon Akame<sup>19</sup>, Augustine B. Arreyetta-Bawak<sup>20</sup>, Oyere T. Ebob<sup>1</sup>, Jonathan A. Metuge<sup>20</sup>, Boris D. Bekono<sup>21</sup>, Mustafa A. Isa<sup>22</sup>, Raphael Onuku<sup>23</sup>, Daniel M. Shadrack<sup>24</sup>, Thommas M. Musyoka<sup>25</sup>, Vaishali M. Patil<sup>26</sup>, Justin J. van der Hooft<sup>27</sup>, Vanderlan da Silva Bolzani<sup>9</sup>, José L. Medina-Franco<sup>15</sup>, Johannes Kirchmair<sup>28</sup>, Tilmann Weber<sup>29</sup>, Özlem Tastan Bishop<sup>25</sup>, Marnix H. Medema<sup>27</sup>, Ludger A. Wessjohann<sup>30,31</sup> and Jutta Ludwig-Müller<sup>3\*</sup>

## Abstract

We report the major conclusions of the online open-access workshop “Computational Applications in Secondary Metabolite Discovery (CAiSMD)” that took place from 08 to 10 March 2021. Invited speakers from academia and industry and about 200 registered participants from five continents (Africa, Asia, Europe, South America, and North America) took part in the workshop. The workshop highlighted the potential applications of computational methodologies in the search for secondary metabolites (SMs) or natural products (NPs) as potential drugs and drug leads. During 3 days, the participants of this online workshop received an overview of modern computer-based approaches for exploring NP discovery in the “omics” age. The invited experts gave keynote lectures, trained participants in hands-on sessions, and held round table discussions. This was followed by oral presentations with much interaction between the speakers and the audience. Selected applicants (early-career scientists) were offered the opportunity to give oral presentations (15 min) and present posters in the form of flash presentations (5 min) upon submission of an abstract. The final program available on the workshop website (<https://caismd.indiayouth.info/>) comprised of 4 keynote lectures (KLs), 12 oral presentations (OPs), 2 round table discussions (RTDs), and 5 hands-on sessions (HSs). This meeting report also references internet resources for computational biology in the area of secondary metabolites that are of use outside of the workshop areas and will constitute a long-term valuable source for the community. The workshop concluded with an online survey form to be completed by speakers and participants for the goal of improving any subsequent editions.

\*Correspondence: fidele.ntie-kang@ubuea.cm;  
jutta.ludwig-mueller@tu-dresden.de

<sup>1</sup> Department of Chemistry, University of Buea, P. O. Box 63, Buea, Cameroon

<sup>3</sup> Institute of Botany, Technische Universität Dresden, Zellescher Weg 20b, 01062 Dresden, Germany

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

**Keywords:** Bioinformatics, Chemoinformatics, Metabolites, Online workshop, Predictions, Web tools

## Introduction

Natural products (NPs) have potential therapeutic uses, either directly as drugs or as lead compounds [1]. The discovery of secondary metabolites (SMs) from bacteria, fungi, and plants as lead compounds for drug discovery purposes by pharmaceutical companies had been slowed down before the last decade, despite their huge representation among compounds approved as drugs. For example, in the area of cancer drug discovery, during the period 1946–1980, 40 out of the 75 approved small molecules by the United States Food and Drug Administration (FDA) were with NPs or NP-derived [2]. During the past decade, SM discovery has been enhanced by the rapid progress in artificial intelligence and its applications [3]. Research in the field of NPs has therefore embraced the need for large-scale analysis of digitized experimental data in the fields of metabolomics, transcriptomics, genomics, often referred to as the “omics” era [4]. This calls for the need for NP chemists to be properly trained in the new “omics” disciplines to be able to tackle the new challenges in the identification of SM, elucidation of their structures, modes of action, and potential toxicities in order to enhance drug discovery from nature.

With the advent of the COVID-19 lockdown associated with travel restrictions and social distancing measures, scientists had to resort to training and sharing of research results via the web [5]. Although online teaching may prove challenging in the sense that it is often difficult to ascertain the level of concentration of the learners and it is not possible to get their immediate response from facial gestures, distance learning has proven to be one of the feasible approaches to ensure that teaching and learning still continue in the midst of the pandemic [6, 7]. Most institutions (mostly in secondary and tertiary education) have resorted to online teaching, while some are still battling with maintaining a minimal amount of face-to-face teaching. This is often for social reasons and to ensure that the educators and learners get feedback from the past lessons and have the possibility to correct exercises and get responses from pressing issues and misunderstood concepts. As a result, modern tools to enhance learning while maintaining barrier measures are in high demand and web conferences have almost completely replaced the traditional scientific conferences and workshops, which have now undergone some hibernation to slow down the spread of the virus [8, 9]. Within the context of the German Academic Exchange Services (DAAD) funding scheme, invited scholars from abroad are encouraged to organize a training event, which could

be in the form of a digitalized lecture (i.e. workshops, conferences, course tutorials, etc.) (<https://www.daad.org/en/find-funding/faculty/visiting-professorship/>). It is within this context that we proposed to organize the online workshop entitled “Computational Applications in Secondary Metabolite Discovery” (CAiSMD).

This virtual workshop introduced participants to modern computer-based approaches and tools for the exploration of the NPs and “omics” world. Most of the tools (software, web servers, databases, etc.), methods and results presented to the participants were recent (dating from 2019 and later). The focus was on bioinformatics, chemoinformatics, NP chemistry, computational drug design, and genomic analysis, with applications in drug discovery. The organizers took the initiative to start identifying, inviting, and corresponding with key experts in the field who could provide inputs in the form of keynote lectures, oral presentations and online hands-on training tutorials. All information regarding deadlines and registration were published on the workshop website from which interested applicants could download an abstract template and make all relevant uploads. The cost-free workshop was conducted in English and open to the entire scientific community. M.Sc. and Ph.D. students, postdoctoral researchers and early-career scientists were the target group of the workshop. Selected participants could submit an abstract indicating if they wish to give 15-min oral presentations.

All sessions with oral presentations and the parallel hands-on sessions (HS) were accessible through Zoom. All digital references are summarized in Table 1 and the final program and hand-out sessions are available on the web site (<https://caismd.indiayouth.info/>). Since the workshop was intended in large part to attract early-career researchers, two formats were included to attract them: parallel hands-on sessions and round table discussions, partially led by early-career post-docs (Additional files 1, 2).

## Workshop contents

### Keynote lectures

Four keynote lectures (KLs) were given during the workshop, each lasting 45 min. The first KL was held by Ludger Wessjohann, who presented an overview of how various informatics methods and tools developed by his group (along with partners) could support the selection of sources and the identification of NPs of relevance [10–12]. The lecture was focused on the chemoinformatic analysis of:

**Table 1** Important web links shared during the CAiSMD workshop

Speaker (lecture)	Group website (web link to tools presented)	References
Ludger A. Wessjohann (KL01) Email: wessjohann@ipb-halle.de	Research group: <a href="http://www.ipb-halle.de/en/research/bioorganic-chemistry/">http://www.ipb-halle.de/en/research/bioorganic-chemistry/</a> Tools: Databases and Tools, e.g. for computational mass spectrometry ( <a href="https://www.ipb-halle.de/en/infrastructure/databases-and-tools/?MP=6-1058">https://www.ipb-halle.de/en/infrastructure/databases-and-tools/?MP=6-1058</a> )	[48–50]
Marnix Medema (KL02) Email: marnix.medema@wur.nl	Research group: <a href="http://marnixmedema.nl/">http://marnixmedema.nl/</a> Tools: the antiSMASH software ( <a href="https://antismash.secondarymetabolites.org/">https://antismash.secondarymetabolites.org/</a> ); the antiSMASH database ( <a href="https://antismash-db.secondarymetabolites.org/">https://antismash-db.secondarymetabolites.org/</a> ); Paired Omics Data Platform ( <a href="https://pairedomicsdata.bioinformatics.nl/">https://pairedomicsdata.bioinformatics.nl/</a> ); the plantiSMASH software ( <a href="http://plantismash.secondarymetabolites.org/">http://plantismash.secondarymetabolites.org/</a> )	[13–15, 51, 52]
Tilman Weber (KL03) Email: tiwe@biosustain.dtu.dk	Research group: <a href="https://www.biosustain.dtu.dk/">https://www.biosustain.dtu.dk/</a> Tools: the antiSMASH software ( <a href="https://antismash.secondarymetabolites.org/">https://antismash.secondarymetabolites.org/</a> ); the antiSMASH database ( <a href="https://antismash-db.secondarymetabolites.org/">https://antismash-db.secondarymetabolites.org/</a> ); CRISPy ( <a href="https://crispy.secondarymetabolites.org/">https://crispy.secondarymetabolites.org/</a> ); the secondary metabolite bioinformatics portal ( <a href="http://www.secondarymetabolites.org/">http://www.secondarymetabolites.org/</a> )	[14, 15, 52, 53]
Özlem TASTAN Bishop (KL04) Email: o.tastanbishop@ru.ac.za Justin J. J. van der Hooft (HS01) Email: justin.vanderhooft@wur.nl	Research group: <a href="https://rubi.ru.ac.za/">https://rubi.ru.ac.za/</a> Tools: SANCDDB compound database ( <a href="https://sancdb.rubi.ru.ac.za/">https://sancdb.rubi.ru.ac.za/</a> ) Research group: <a href="https://www.wur.nl/en/Persons/Justin-dr.-JJJ-Justin-van-der-Hooft.htm">https://www.wur.nl/en/Persons/Justin-dr.-JJJ-Justin-van-der-Hooft.htm</a> Tools: GNPS ( <a href="https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp">https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp</a> ); MS2LDA ( <a href="http://www.ms2lda.org/">http://www.ms2lda.org/</a> ); MAGMa ( <a href="https://www.emetabolomics.org/">https://www.emetabolomics.org/</a> ); Paired Omics Data Platform ( <a href="https://pairedomicsdata.bioinformatics.nl/">https://pairedomicsdata.bioinformatics.nl/</a> )	[35, 54] [34, 51, 52, 55, 56]
Daniel M Shadrack (HS02) Email: dmssjut@gmail.com Thommas Musyoka (HS02) Email: mutemibiochemistry@gmail.com Fidele Ntie-Kang (HS02) Email: fidele.ntie-kang@ubuea.cm Yannick Djoumbou-Feunang (HS03) Email: yannick.djoumboufeunang@corteva.com Conrad Stork (HS04) Email: stork@zbh.uni-hamburg.de	Research group: Tools: none Research group: same as KL04 Tools: same as KL04 Research group: <a href="https://ntiekfidele.jimdofree.com/">https://ntiekfidele.jimdofree.com/</a> Tools: ANPDB compound database ( <a href="http://african-compounds.org/anpdb/">http://african-compounds.org/anpdb/</a> ) Research group: Tools: BioTransformer ( <a href="http://biotransformer.ca/">http://biotransformer.ca/</a> ) Research group: <a href="https://www.zbh.uni-hamburg.de/personen/acm/cstork.html">https://www.zbh.uni-hamburg.de/personen/acm/cstork.html</a> Tools: NERDD web server ( <a href="https://nerdd.univie.ac.at/">https://nerdd.univie.ac.at/</a> ); FAME3 ( <a href="https://nerdd.univie.ac.at/fame3/">https://nerdd.univie.ac.at/fame3/</a> ); GLORY ( <a href="https://nerdd.univie.ac.at/glory/">https://nerdd.univie.ac.at/glory/</a> ); GLORYx ( <a href="https://nerdd.univie.ac.at/gloryx/">https://nerdd.univie.ac.at/gloryx/</a> ); Hit Dexter 2.0 ( <a href="https://nerdd.univie.ac.at/hitdexter/">https://nerdd.univie.ac.at/hitdexter/</a> ); NP-Scout ( <a href="https://nerdd.univie.ac.at/npscout/">https://nerdd.univie.ac.at/npscout/</a> ); Skin Doctor CP ( <a href="https://nerdd.univie.ac.at/skinDoctor/">https://nerdd.univie.ac.at/skinDoctor/</a> )	[36, 37] [26] [17–23]
Neann Mathai (HS04) Email: neann.mathai@uib.no Maria Sorokina (HS05) Email: maria.sorokina@uni-jena.de	Research group: <a href="https://www.uib.no/en/persons/Neann.Sarah.Mathai">https://www.uib.no/en/persons/Neann.Sarah.Mathai</a> Tools: Same as above Research group: <a href="https://cheminf.uni-jena.de/members/aria-sorokina/">https://cheminf.uni-jena.de/members/aria-sorokina/</a> Tools: CDK ( <a href="https://cdk.github.io/">https://cdk.github.io/</a> ); COCONUT compound database ( <a href="https://coconut.naturalproducts.net/">https://coconut.naturalproducts.net/</a> ); NP-likeness score web application ( <a href="https://naples.naturalproducts.net/">https://naples.naturalproducts.net/</a> ); Sugar Removal Utility ( <a href="https://sugar.naturalproducts.net/">https://sugar.naturalproducts.net/</a> )	[38–43]
Kiran K. Telukunta (OP01) Email: kiran.telukunta@indiayouth.info Paul F. Zierrep (OP02) Email: paul.zierrep@gmail.com	Research group: Tools: galaxy tutorials ( <a href="https://galaxyproject.org/learn/">https://galaxyproject.org/learn/</a> ) Research group: <a href="http://www.pharmbioinf.uni-freiburg.de/main/members/ziereppaul/">http://www.pharmbioinf.uni-freiburg.de/main/members/ziereppaul/</a> Tools: sempi 2.0 web server ( <a href="http://sempi.pharmazie.uni-freiburg.de/">http://sempi.pharmazie.uni-freiburg.de/</a> ); Strep-tomeDB compound database ( <a href="http://thymin.pharmazie.uni-freiburg.de/streptomedb/">http://thymin.pharmazie.uni-freiburg.de/streptomedb/</a> ); the ANPDB compound database ( <a href="http://african-compounds.org/anpdb/">http://african-compounds.org/anpdb/</a> )	[57] [22, 36, 37, 58]
Victor C. Osamor (OP03) Email: victor.osamor@covenantuniversity.edu.ng; vcosamor@gmail.com Marilia Valli (OP04) Email: mariliava@gmail.com Vanderlan da Silva Bolzani (OP04)	Research group: <a href="https://covenantuniversity.edu.ng/Profiles/Osamor-Victor-Chukwudi#.YGYag1UzblU">https://covenantuniversity.edu.ng/Profiles/Osamor-Victor-Chukwudi#.YGYag1UzblU</a> Tools: OsamorSoft (not available) Research group: <a href="https://nubbe.iq.unesp.br/portal/index.html">https://nubbe.iq.unesp.br/portal/index.html</a> Tools: NuBBE database ( <a href="https://nubbe.iq.unesp.br/portal/nubbe-search.html">https://nubbe.iq.unesp.br/portal/nubbe-search.html</a> ) Research group: same as above	[23] [59, 60]

**Table 1** (continued)

Speaker (lecture)	Group website (web link to tools presented)	References
Email: vanderlan.bolzani@unesp.br Samuel Egjeyeh (OP05)	Tools: same as above Research group:	
Email: segjeyeh@uwc.ac.za Miquel Duran-Frigola (OP06)	Tools: none Research group: <a href="https://ersilia.io/">https://ersilia.io/</a>	
Email: miquel@ersilia.io Yannick Djoumbou-Feunang (OP07)	Tools: Chemical Checker ( <a href="https://pypi.org/project/chemicalchecker/">https://pypi.org/project/chemicalchecker/</a> ) Research group: same as HS03	[61]
Email: yannick.djoumboufeunang@corteva.com Ana L. Chávez-Hernández (OP08)	Tools: same as HS03 Research group: <a href="https://www.difacqum.com/">https://www.difacqum.com/</a>	
Email: anachavez3026@gmail.com  José L. Medina-Franco (OP08)	Tools: Epigenetic Target Profiler web server ( <a href="http://www.epigenetictargetprofiler.com/">http://www.epigenetictargetprofiler.com/</a> ); BIOFACQUIM natural products database and other tools, e.g. Platform for Unified Molecular Analysis (PUMA), Activity Landscape Plotter, etc. ( <a href="https://www.difacqum.com/d-tools/">https://www.difacqum.com/d-tools/</a> ) Research group: same as above	[62, 63]
Email: jose.medina.franco@gmail.com Johannes Kirchmair (OP09)	Tools: same as above Research group: <a href="https://comp3d.univie.ac.at/the-comp3d-team/johannes-kirchmair/">https://comp3d.univie.ac.at/the-comp3d-team/johannes-kirchmair/</a>	
Email: johannes.kirchmair@univie.ac.at Maria Sorokina (OP10)	Tools: same as HS04 Research group: same as HS05	
Email: maria.sorokina@uni-jena.de Vaishali M. Patil (OP11)	Tools: same as HS05 Research group:	
Email: vaishuwise@gmail.com Romuald Tematio Fouedjou (OP12)	Tools: none Research group:	
Email: tematioromy@yahoo.fr	Tools: none	

1. Large databases and text corpuses, and
2. The role of metabolic profiling in future bioactive compound discovery, going beyond classical isolation processes.

An example included in the lecture was a survey of the flora of Java (Indonesia), metabolic profiling of various medicinal plants, with an emphasis on *Hypericum* spp. (St. John's wort).

At the beginning of day 2, the participants followed the second keynote lecture by Marnix Medema, who highlighted the ongoing work in his research group. His group is developing and using computational methods to identify plant-based, fungal, and bacterial molecules of ecological and clinical importance and developing approaches to assess and predict the biological activities of specialized metabolites to accelerate NP discovery by focusing efforts on the most promising candidates. In his keynote, Medema showed the importance of computational methodologies in the study of microbe-microbe and host-microbe interactions in human, plant and animal microbiomes. Specifically, he discussed the use of computational approaches to investigate biosynthetic diversity across large numbers of genomes, and integrative genome/metabolome mining to link gene clusters to molecules facilitated by novel community-based efforts

such as the Paired Omics Data Platform (<https://pairedomicsdata.bioinformatics.nl>) [13].

The third KL took place at the end of day 2, during the session entitled “looking into the future”, after several oral presentations (OPs). Tilmann Weber presented the latest version (v6) of the antiSMASH genome mining platform (<https://antismash.secondarymetabolites.org>) [14], which has been under development since 2011, coordinated by his group and that of Marnix Medema. The version 6 of antiSMASH includes an improved user interface, new detection modules, a new cluster comparison tool, and many internal optimizations. These were presented as useful tools for the easy analysis of genomic sequences for the presence of secondary metabolite biosynthetic gene clusters (BGCs) in bacteria and fungi. Additionally, the antiSMASH database (<https://antismash-db.secondarymetabolites.org/>) [15] was presented as a user-friendly application that allows users to browse and query pre-computed antiSMASH v5 annotations. The database contains information on 147,517 high-quality BGC regions from 388 archaeal, 25,236 bacterial and 177 fungal genomes. It was highlighted that these basic genome mining technologies build the foundations of further in silico studies towards a more comprehensive “Genome Analytics” platform, which could be used to streamline NP discovery and characterization efforts in the future.

The fourth KL was given by Özlem Tastan Bishop on day 3. The presenter showed some early drug discovery research experiences within the context of Africa. The main interests include the identification of novel and alternative drug targeting sites (i.e. allosteric sites) and of hit compounds for communicable and non-communicable diseases, which have recently been published by her group [16–21]. Her group had shown interest to understand the effects of nonsynonymous single nucleotide variations (nsSNVs) on protein structure and function, in order to (i) assess the reasons behind many inherited diseases, (ii) uncover the association to drug resistance mechanisms, and (iii) link to drug sensitivity issues in certain populations for precision medicine development; and many other applications. The lecturer also argued that an understanding of the underlying resistance mechanism due to variations at the molecular level is essential and can lead either to modifications of currently approved drugs to get more effective ones or to the design of new inhibitors that overcome resistance mutations. The lecture concluded with recent work on understanding the underlying drug resistance mechanisms and identification of allosteric modulators as an alternative to orthosteric drugs.

#### Oral presentations

There were 12 scheduled OPs. The lecture topics included presentations of both novel computational methodologies as well as recent results, e.g., methods for clustering of specialized metabolites and the introduction of a large integrated and open database for NPs (<https://coconut.naturalproducts.net>) and the most recent version of the NuBBE natural products database from Brazil. Other speakers presented web servers for prediction of metabolites from gene cluster data, e.g., the SeMPI 2.0 web server (<http://sempi.pharmazie.uni-freiburg.de/>) presented by Paul Zierep for polyketide synthase (PKS) and non-ribosomal peptide synthase (NRPS) prediction by combining with metabolite screening in natural product databases [22] or drug discovery platforms, e.g. that of the University of West Cape (South Africa) presented by Samuel A. Egieyeh. Other methods presented were those implemented in software, e.g. the OsamorSoft spreadsheets, the tool developed in the group of Victor C. Osamor for clustering genomic data [23], with potentially useful applications in plant-based SMs as lead compounds in two databases from African medicinal plants, with a focus on the NANPDB and EANPDB databases (see Table 1).

The afternoon and early evening sessions consisted of 5 core cheminformatics lectures. In the quest to predict the biological activities of non-characterized NPs, Miquel Duran-Frigola presented the tool Chemical

Checker (CC), which included a collection of deep neural networks capable of inferring bioactivity signatures for any compound of interest, even when little or no experimental information is available for them [24]. The speaker also showed how inferred bioactivity signatures are useful to navigate the chemical space in a biologically relevant manner, unveiling higher-order organization in NP collections. The lecturer also used an implementation of a battery of signature-activity relationship models to show that this resulted in a substantial improvement in predictive performance, with respect to chemistry-based classifiers, across a series of biophysics and physiology activity prediction benchmarks [25]. It could, therefore, be concluded that from the CC, large-scale inference of bioactivity profiles can set the basis for automated annotation of compound collections, including drugs, metabolites and NPs.

Yannick Djoumbou-Feunang highlighted how artificial intelligence in the form of machine learning could be useful for secondary metabolite prediction. *In silico* metabolism prediction tools provide a unique perspective to studying the chemical exposome, and how its changes affect the environment. Classical applications of such tools include, but are not limited to metabolite discovery, environmental fate prediction, ADMET profiling, and molecular design. Several approaches and methods to address the prediction of secondary metabolites have been described, and implemented in a comprehensive list of tools that include expert-, machine learning-, and quantum mechanics (QM)-based systems, or hybrids thereof. However, the speaker showed that in spite of the numerous reported successes, many limitations still hamper the wide adoption of those tools. In his presentation, he described the impact of artificial intelligence in the development of secondary metabolite prediction systems, along with the most commonly implemented approaches. He then showed examples of the application of *in silico* metabolism prediction tools, such as BioTransformer which he developed [26], in the identification of secondary metabolites. This tool could be useful for identifying the plausible metabolites of a compound, including NPs. He concluded by showing some of the prevalent limitations that hamper the widespread adoption of such tools and propose solutions.

Next, Ana L. Chávez-Hernández, a Ph.D. student from José L. Medina-Franco's team in Mexico, presented a fragment library of NPs and compound databases for drug discovery, laying emphasis on the fragment libraries were generated from the recently published COLleCtion of Open NatUral producTs (COCONUT) [27] and other reference-data sets such as food chemical compounds (FooDB) [28], including compounds from a focused library from the Chemical Abstract Service (CAS) and

inhibitors of the main protease of SARS-CoV-2 (3CLP). The fragment libraries generated from COCONUT, the library focused on COVID-19 research and other reference databases are publicly available in [29].

Johannes Kirchmair provided a succinct overview of the state-of-the-art in computational target prediction, with a focus on NPs research. He discussed the scope and limitations of 2D and 3D similarity-based methods, network-based approaches, machine learning models and docking approaches. Kirchmair talked about guidelines on how to make the best use of *in silico* models and understand the reliability of predictions.

Maria Sorokina presented the recently available MongoDB [30], a document-based noSQL database management system particularly suitable for storing very diverse and sparse data on NPs. Due to easy data querying and crossing, she highlighted that this database type is rapidly gaining popularity in the cheminformatics community, as more and more chemical interfaces for it are developed to enable similarity and substructure searches. She concluded her talk by showing the example of the COCONUT natural products database [31] as an illustration of a noSQL database. During the OP of Romuald Tematio Fouedjou which took place during the morning session of day 3 (dedicated to young investigators), a virtual procedure aimed at the identification of NPs with a potential affinity towards the SARS-CoV-2 main protease and spike protein based on compounds from Cameroonian medicinal plants of the Asteraceae was presented. Although the results were somewhat preliminary, the goal was to identify inhibitors that could facilitate the development of potential anti-COVID-19 drug lead compounds from Cameroonian medicinal plants.

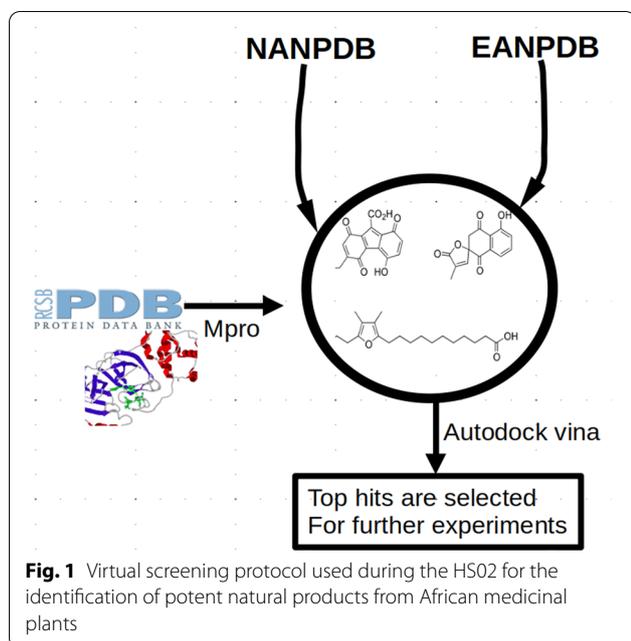
#### Hands-on sessions

Day 3 started with the parallel running of five hands-on sessions (HS). Each lasted 90 min and it was possible to switch between sessions. These sessions were received very well by the participants since they focused on direct training with selected databases and tools.

During HS01, the respective participants followed a hands-on tutorial on mining the plant specialized metabolome with help of mass spectrometry data in the Global Natural Product Social Molecular Networking (GNPS) environment (<https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp>) [32]. In particular, library matching and molecular networking were touched upon as well as how to study the occurrence of particular metabolites across a plant dataset when quantification information is available; for example, across genera or clades. After the main motivation to develop metabolome mining tools and a little theory, the participants got the chance to inspect the reliability of the library matches found in

a recently published work studying >70 Rhamnaceae plants from various genera and two main clades [33]. This was followed by the study of a selected number of molecular families to learn more about how molecular networking can group structurally related metabolites to propagate structural annotations within molecular families and to facilitate their joint analysis. Finally, the inclusion of quantification data into the analysis through Feature-based Molecular Networking [34] was exploited by inspecting the abundance of a number of triterpenoid and flavonoid metabolite features from across genera and clades. Even though some of these metabolites seemed highly related based on their mass fragmentation data, their abundance patterns were quite different. The hands-on session, thus, offered a quick way of reproducing and validating the results as presented in the paper [33]. The session ended with an outlook on exciting future developments in chemical compound class annotation and mass spectral similarity metrics.

The second hands-on session focused on virtual screening for the identification of bioactive SMs from NP databases, with a focus on NPs from Africa. The first presenter (Thommas M. Musyoka) introduced the South African Natural Compounds Database (SANCDB—<https://sancdb.rubi.ru.ac.za/>) [35], a collection of 1012 compounds derived from South African natural sources. Since its inception in 2015, the database has been used for various machine learning and *in silico* virtual drug screening studies with a recent study identifying several potential hits against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As part of a recent update, a unique feature incorporating the compound dataset analogs from two leading commercial databases (Molport—<https://www.molport.com/> and Mcule—<https://mcule.com/>) was included. The feature not only allows users to explore a larger chemical space during screening but also enables them to seamlessly purchase compounds for their biological studies. Participants were introduced to the database by the speaker Fidele Ntie-Kang, who emphasized how they can obtain compounds in different chemical formats for both their virtual screening and biological studies. The second part of the session (approximately 20 min) focused on NPs databases originating from the regions of Northern [36] and East Africa [37] (<http://african-compounds.org/anpdb/>). The participants explored the web tools that enhance the search of the databases, including similarity and substructure searching for privileged scaffolds. Afterward they were introduced to the NP databases from African sources, their contents, compound classes and potential for lead compound discovery. During the last session (about 50 min), Daniel M. Shadrack introduced the participants to state-of-the-art computational techniques used in lead



compound identification from electronic databases, e.g. molecular docking, and pharmacophore-based searching of the NP databases. In this section, participants were introduced to the approaches used to perform in silico screening of libraries containing natural products against the SARS-CoV-2 main protease by screening the NANPDB and EANPDB datasets against the target using the Autodock tool (see Fig. 1). The participants were also briefly introduced to other sophisticated tools like molecular dynamics and metadynamics, just on the fly, with the goal of learning how to perform virtual screening from large libraries (focusing on natural products libraries from African sources). Using this as an example, a similar approach could be used in any other project of interest, involving a different drug target.

HS03 was conducted by Yannick Djoumbou-Feunang. The first part of this session was focused on describing BioTransformer [26], an open-source software tool, and freely accessible server for the prediction of human cytochrome P450-catalyzed metabolism, human gut microbial degradation, human phase-II metabolism, human promiscuous metabolism, and environmental microbial degradation. Additionally, BioTransformer assists in metabolite identification, and metabolic pathway prediction. In the second part, an assessment of BioTransformer's performance in the prediction of metabolism for diverse sets of molecules, including but not limited to pharmaceuticals, pesticides, and phytochemicals was presented. Overall, BioTransformer was shown to

achieve moderately high precision ( $>0.46$ ), and higher recall ( $>0.84$ ) when predicting human metabolism of drugs, lipids, and phytochemicals, compared to two commercially available tools. On the other hand, the overall precision ( $\sim 0.3$ ) and recall ( $\sim 0.6$ ) achieved for the metabolism prediction of agrochemicals suggest that improvements are needed to cover more chemicals as well as biological species that are relevant to agrosiences. In the third part of this presentation, an illustration of a few examples of its application as demonstrated by various published scientific studies was provided. Furthermore, the presenter shared future perspectives for this open-source project and described how it could significantly benefit the exposure science and regulatory communities.

During the fourth HS jointly coordinated by Conrad Stork and Neann Mathai, both Ph.D. students of Johannes Kirchmair, the New E-Resource for Drug Discovery (NERDD) web service [38] was introduced to the audience. NERDD, available at <https://nerdd.univie.ac.at/>, provides an intuitive user interface to six cheminformatics tools: FAME3 [39] for site-of-metabolism prediction, GLORY [40] and GLORYx [41] for metabolite (structure) prediction, Hit Dexter 2.0 [42] for the prediction of frequent hitters in biological assays, NP-Scout [43] for the prediction of natural products, and Skin Doctor CP [44] for the prediction of the skin sensitization potential of small molecules. The participants were guided through the usage of each of the tools starting from the query upload to the result interpretation. For each of the tools, an exercise was given to show a range of use cases which were discussed cooperatively with the listeners.

The fifth parallel session was led by Maria Sorokina and focused on the Chemistry Development Kit (CDK) [45]. The CDK is one of the main programming toolkits for processing and analyzing chemical information. The only prerequisite for attending this workshop was some coding experience. The CDK is available as modular Java libraries, easy to use, open and free, and available to download at <https://cdk.github.io/>. It can also be easily integrated with both Maven and Gradle. During this hands-on, the key CDK concepts were presented, such as molecule representation and manipulation, diverse fingerprints, and molecular descriptors. Examples of the usage of CDK for natural product discovery and analysis, like the NP-likeness scorer [46] and the Sugar Removal Utility [47], were also presented. All code examples presented during this workshop were provided (Table 1). The only prerequisite for attending this workshop was some coding experience.

### Round table discussions

There were two round table discussions (RTDs). During RTD01 (following the session on bioinformatics applications which addressed issues like clustering in metabolite discovery and presentations of several web servers like SeMPI 2.0 and clustering tools Osamor-Soft), several interesting questions came up, e.g., which tools would be the best to address the increasing volume of data on secondary metabolites, genomic sequences, transcriptomes, etc. To address this question, the panelists pointed to a more multidisciplinary approach that integrates the various “omics” datasets. Besides, questions were raised regarding the future of cheminformatics for biodiversity research in Brazil, as the morning speaker (Marilia Valli) had mentioned that many of the biodiversity hotspots were not well represented in their collected NuBBE database. To address this, the speaker mentioned that although there is still a majority of species in Brazil to be studied, the published works are being added to NuBBE Database. During the RTD02 chaired by early career researchers, Serge A. T. Fobofou first of all appreciated the organizers for the effort put together to bring together top researchers to such a workshop and for making the workshop freely accessible to all, thus saving a huge financial cost. Feedback from the parallel session was received during the first part of the round table discussion that followed that session (RTD02). In general, the participants agreed with the current format of online presentations, but would have preferred to have a workshop with physical presence. The chair further invited participants to ask specific questions to the individual speakers before making suggestions for the future. The questions from the participants pointed to the fact that the workshop contents were received with a lot of enthusiasm, suggesting that a compendium of computational tools that were introduced during the workshop be compiled and made available to all participants. This has been made available in Table 1.

### Conclusions

SMs play important roles in agricultural, cosmetic and pharmaceutical industries and the mastery of the use of computational methodologies for their investigation is urgently needed in the scientific community. The importance of this workshop is that it was able to bring together 195 registered participants and 24 experts from around the world, and that the participants would benefit from 3 days of intensive training free of cost. Basically, the attendees were drilled on state-of-the-art in silico methodologies, algorithms and tools that could be useful in the rapid discovery of NPs for small molecule drug discovery. A compendium of in silico tools to enhance

NP dereplication, lead discovery and de-novo design was made available to all participants. The workshop lectures and materials shown and worked with during the hands-on sessions have been made available for download from the workshop website. Participants from low income countries who, otherwise, would not be able to attend such a high profiled meeting could do so at no cost and have access to the software and servers that would enhance their discovery of drugs from NP-based leads. It is hoped that these would serve as a foundation for early-career researchers working or starting off their studies in this field. Besides, the responses from the feedback survey will serve the purpose of improving subsequent similar events. The workshop ended with a round table discussion led by early-career researchers to summarize the key events of the meeting.

### Abbreviations

ANPDB: African Natural Products Database; BGC: Biosynthetic gene cluster; CAS: Chemical Abstracts Service; CDK: Chemistry Development Kit; GNPS: The Global Natural Product Social Molecular Networking; HS: Hands-on session; KL: Keynote lecture; MAGMa: An online application for the automatic chemical annotation of accurate multistage MS<sup>n</sup> spectral data; MS2LDA: A tool that decomposes molecular fragmentation data derived from large metabolomics experiments into annotated Mass2Motifs; NERDD: New E-Resource for Drug Discovery; NP: Natural product; NRPS: Nonribosomal Peptide-Synthetase; nsSNVs: Nonsynonymous Single Nucleotide Variations; OP: Oral presentation; PKS: Polyketide Synthase; RTD: Round table discussion; SANCDB: South African Natural Compounds Database.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-021-00546-8>.

**Additional file 1.** Workshop statistics.

**Additional file 2.** CAISMD Feedback form.

### Acknowledgements

The Tarunavadaanenasaha Muktbharatonnayana Samstha Foundation (TMS Foundation) is acknowledged for supporting the workshop for its online presence and facilitating the participants to find the content about CAISMD by hosting the information about speakers and events. The workshop organizers also acknowledge some technical support from the IT team of the Technische Universität Dresden, Germany.

### Supplementary information

The workshop slides and materials for the hands-on sessions are available for free download from the website (<https://indiayouth.info/index.php/caismd/caismd-downloads>). Additional useful Workshop and Slides of Post-Workshop Feedback Survey Responses are also uploaded.

### Authors' contributions

FN-K and JL-M conceived the workshop and invited speakers, edited and approved the submitted abstracts, coordinated the organizational aspects of the workshop, and chaired some sessions. KKT contributed in developing the submission portal and workshop website and also gave an oral presentation. All other co-authors either wrote parts of the report, chaired a session, or participated as speakers and as interactive participants by providing feedback contributions. All authors read and approved the final manuscript.

## Funding

Financial support is acknowledged from the German Academic Exchange Services (DAAD) and from Technische Universität Dresden for covering part of the conference cost.

## Declarations

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Department of Chemistry, University of Buea, P. O. Box 63, Buea, Cameroon. <sup>2</sup>Institute of Pharmacy, Martin-Luther University of Halle-Wittenberg, Kurt-Mothes-Str. 3, 06120 Halle, Germany. <sup>3</sup>Institute of Botany, Technische Universität Dresden, Zellescher Weg 20b, 01062 Dresden, Germany. <sup>4</sup>Tarunavadaanenasaha Muktbharatonnayana Samstha Foundation, Hyderabad, India. <sup>5</sup>Institute of Pharmaceutical Biology, Technische Universität Braunschweig, Mendelssohnstrasse 1, 38106 Braunschweig, Germany. <sup>6</sup>Department of Computer and Information Sciences, Colege of Science and Technology, Covenant University, Km. 10 Idiroko Rd, Ogun, Ota, Nigeria. <sup>7</sup>School of Pharmacy, University of the Western Cape, Cape Town 7535, South Africa. <sup>8</sup>South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town 7535, South Africa. <sup>9</sup>Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, Sao Paulo State University-UNESP, Araraquara, Brazil. <sup>10</sup>Corteva Agriscience, Indianapolis, IN, USA. <sup>11</sup>Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University, Jena, Germany. <sup>12</sup>Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany. <sup>13</sup>Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, 5020 Bergen, Norway. <sup>14</sup>Pharmaceutical Bioinformatics, Albert-Ludwigs-University, Freiburg, Germany. <sup>15</sup>DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico. <sup>16</sup>Ersilia Open Source Initiative, Cambridge, UK. <sup>17</sup>Joint IRB-BSC-CRG Programme in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain. <sup>18</sup>Department of Chemistry, University of Dschang, Dschang, Cameroon. <sup>19</sup>Department of Immunology, School of Health Sciences, Catholic University of Central Africa, BP 7871, Yaoundé, Cameroon. <sup>20</sup>Department of Biochemistry and Molecular Biology, University of Buea, P. O. Box 63, Buea, Cameroon. <sup>21</sup>Department of Physics, Ecole Normale Supérieure, University of Yaoundé I, BP. 47, Yaoundé, Cameroon. <sup>22</sup>Bioinformatics and Computational Biology Lab, Department of Microbiology, Faculty of Sciences, University of Maiduguri, P.M.B. 1069, Maiduguri, Borno State, Nigeria. <sup>23</sup>Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria Nsukka, Nsukka, Nigeria. <sup>24</sup>Department of Chemistry, St. John's University of Tanzania, P. O. Box 47, Dodoma, Tanzania. <sup>25</sup>Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Makhanda 6140, South Africa. <sup>26</sup>Computer Aided Drug Design Lab, KIET Group of Institutions, Delhi-NCR, Ghaziabad 201206, India. <sup>27</sup>Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. <sup>28</sup>Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, 1090 Vienna, Austria. <sup>29</sup>The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark. <sup>30</sup>Department of Bioorganic Chemistry, Leibniz Institute of Plant Biochemistry (IPB), Weinberg 3, 06120 Halle (Saale), Germany. <sup>31</sup>German Centre for Integrative Biodiversity Research (iDiv), Puschstraße 4, 04103 Leipzig, Germany.

Received: 22 July 2021 Accepted: 23 August 2021

Published online: 06 September 2021

## References

- Khan N, Chen X, Geiger JD (2021) Possible therapeutic use of natural compounds against COVID-19. *J Cell Signal* 2:63–79
- Newman DJ, Cragg GM (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod* 83:770–803
- Harvey A, Edrada-Ebel R, Quinn R (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* 14:111–129
- van Santen JA, Kautsar SA, Medema MH, Lington RG (2021) Microbial natural product databases: moving forward in the multi-omics era. *Nat Prod Rep* 38:264–278
- Naumann E, Möhring K, Reifenscheid M, Wenz A, Rettig T, Lehrer R, Krieger U, Juhl S, Friedel S, Fikel M, Cornesse C, Blom AG (2020) COVID-19 policies in Germany and their social, political, and psychological consequences. *Eur Policy Anal* 6:191–202
- Trindade AR, Carmo H, Bidarra J (2020) Current developments and best practice in open and distance learning. *Int Rev Res Open Distrib Learn* 1:1–25
- Valentine D. Distance learning: promises, problems, and possibilities. University of Oklahoma. <https://www.westga.edu/~distance/ojlda/fall53/valentine53.html>. Accessed 30 Mar 2021.
- Kerres M (2020) Against all odds: education in Germany coping with Covid-19. *Postdigit Sci Educ* 2:690–694
- Unger S, Meiran WR (2020) Student attitudes towards online education during the COVID-19 viral outbreak of 2020: distance learning in a time of social distance. *Int J Technol Educ Sci* 4:256–266
- Feiner A, Pitra N, Matthews P, Pillen K, Wessjohann LA, Riewe D (2021) Downy mildew resistance is genetically mediated by prophylactic production of phenylpropanoids in hop. *Plant Cell Environ* 44:323–338
- Michels B, Franke K, Weiglein A, Sultani H, Gerber B, Wessjohann LA (2020) Rewarding compounds identified from the medicinal plant *Rhodiola rosea*. *J Exp Biol* 223:jeb223982
- Holzmeyer L, Hartig AK, Franke K, Brandt W, Muellner-Riehl AN, Wessjohann LA, Schnitzler J (2020) Evaluation of plant sources for anti-infective lead compound discovery by correlating phylogenetic, spatial, and bioactivity data. *Proc Natl Acad Sci USA* 117:12444–12451
- Schorn MA, Verhoeven S, Ridder L, Huber F, Acharya DD, Aksenov AA, Aleti G, Moghaddam JA, Aron AT, Aziz S, Bauermeister A, Bauman KD, Baunach M, Beemelmans C, Beman JM, Berlanga-Clavero MV, Blacutt AA, Bode HB, Boullie A, Brejnrod A, Bugni TS, Calteau A, Cao L, Carrión VJ, Castelo-Branco R, Chanana S, Chase AB, Chevrette MG, Costa-Lotufo LV, Crawford JM, Currie CR, Cuyppers B, Dang T, de Rond T, Demko AM, Dittmann E, Du C, Drozd C, Dujardin JC, Dutton RJ, Edlund A, Fewer DP, Garg N, Gauglitz JM, Gentry EC, Gerwick L, Glukhov E, Gross H, Gugger M, Guillén Matus DG, Helfrich EJN, Hempel BF, Hur JS, Iorio M, Jensen PR, Kang KB, Kayser L, Kelleher NL, Kim CS, Kim KH, Koester I, König GM, Leao T, Lee SR, Lee YY, Li X, Little JC, Maloney KN, Männle D, Martin HC, McAvoy AC, Metcalf WW, Mohimani H, Molina-Santiago C, Moore BS, Mullenwey MW, Muskat M, Nothias LF, O'Neill EC, Parkinson EI, Petras D, Piel J, Pierce EC, Pires K, Reher R, Romero D, Roper MC, Rust M, Saad H, Saenz C, Sanchez LM, Sørensen SJ, Sosio M, Süßmuth RD, Sweeney D, Tahlan K, Thomson RJ, Tobias NJ, Trindade-Silva AE, van Wezel GP, Wang M, Weldon KC, Zhang F, Ziemert N, Duncan KR, Crüsemann M, Rogers S, Dorrestein PC, Medema MH, van der Hoof JJJ (2021) A community resource for paired genomic and metabolomic data mining. *Nat Chem Biol* 17(4):363–368
- Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 47:W81–W87
- Blin K, Shaw S, Kautsar SA, Medema MH, Weber T (2021) The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res* 49:D639–D643
- Sheik Amamuddy O, Verkhivker GM, Tastan Bishop Ö (2020) Impact of early pandemic stage mutations on molecular dynamics of SARS-CoV-2 M<sup>pro</sup>. *J Chem Inf Model* 60:5080–5102
- Sheik Amamuddy O, Musyoka TM, Boateng RA, Zabo S, Tastan Bishop Ö (2020) Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of *Mycobacterium tuberculosis* pyrazinamidase. *Comput Struct Biotechnol J* 18:1103–1120
- Sheik Amamuddy O, Veldman W, Manyumwa C, Khairallah A, Agajanian S, Oluyemi O, Verkhivker G, Tastan Bishop Ö (2020) Integrated computational approaches and tools for allosteric drug discovery. *Int J Mol Sci* 21:847
- Amusengeri A, Astl L, Lobb K, Verkhivker GM, Tastan Bishop Ö (2019) Establishing computational approaches towards identifying malarial

- allosteric modulators: a case study of *Plasmodium falciparum* Hsp70s. *Int J Mol Sci* 20:5574
20. Penkler DL, Tastan Bishop Ö (2019) Modulation of human Hsp90a conformational dynamics by allosteric ligand interaction at the C-terminal domain. *Sci Rep* 9:1600
  21. Amusengeri A, Tastan Bishop Ö (2019) Discorhabdin N, a South African natural compound, for Hsp72 and Hsc70 allosteric modulation: combined study of molecular modeling and dynamic residue network analysis. *Molecules* 24:188
  22. Zierep PF, Ceci AT, Dobrusin I, Rockwell-Kollmann SC, Günther S (2020) SeMPI 2.0-A Web Server for PKS and NRPS predictions combined with metabolite screening in natural product databases. *Metabolites* 11:13
  23. Osamor IP, Osamor VC (2020) OsamorSoft: clustering index for comparison and quality validation in high throughput dataset. *J Big Data* 7:48
  24. Bertoni M, Duran-Frigola M, Badia-i-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llgero A, de Herreros AG, Aloy P (2021) Bioactivity descriptors for uncharacterized compounds. *BiorXiv*. <https://doi.org/10.1101/2020.07.21.214197>
  25. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2017) MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 9:513–530
  26. Djombou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform* 11:2
  27. Sorokina M, Steinbeck C (2020) Review on natural products databases: where to find data in 2020. *J Cheminform* 12:20
  28. The Metabolomics Innovation Centre. The Metabolomics Innovation Centre: FoodB (version 1). <https://foodb.ca/>. Accessed 28 Mar 2021.
  29. Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL (2020) Fragment library of natural products and compound databases for drug discovery. *Biomolecules* 10:1518
  30. mongoDB: the database for modern applications. <https://www.mongodb.com/de>.
  31. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT online: collection of open natural products database. *J Cheminform* 13:2
  32. Aron AT, Gentry EC, McPhail KL, Nothias LF, Nothias-Esposito M, Bouslimani A, Petras D, Gauglitz JM, Sikora N, Vargas F, van der Hooft JJJ, Ernst M, Kang KB, Aceves CM, Caraballo-Rodríguez AM, Koester I, Weldon KC, Bertrand S, Roullier C, Sun K, Tehan RM, Boya PCA, Christian MH, Gutiérrez M, Ulloa AM, Tejada Mora JA, Mojica-Flores R, Lakey-Beitia J, Vásquez-Chaves V, Zhang Y, Calderón AI, Tayler N, Keyzers RA, Tugizimana F, Ndlovu N, Aksenov AA, Jarmusch AK, Schmid R, Truman AW, Bandeira N, Wang M, Dorrestein PC (2020) Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat Protoc* 15:1954–1991
  33. Kang KB, Ernst M, van der Hooft JJJ, da Silva RR, Park J, Medema MH, Sung SH, Dorrestein PC (2019) Comprehensive mass spectrometry-guided phenotyping of plant specialized metabolites reveals metabolic diversity in the cosmopolitan plant family Rhamnaceae. *Plant J* 98:1134–1144
  34. Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protasyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard PM, Barsch A, Cachet X, Caraballo-Rodríguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kamenik Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M, Martin HC, McCall LI, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyné O, Neumann S, Neuweyer H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Böcker S, Alexandrov T, Bandeira N, Wang M, Dorrestein PC (2020) Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* 17:905–908
  35. Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, Tastan Bishop Ö (2015) SANCDB: a South African natural compound database. *J Cheminform* 7:29
  36. Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, Moumbock AFA, Malange YI, Njume LE, Yong JN, Sippl W, Günther S (2017) NANPDB: a resource for natural products from Northern African sources. *J Nat Prod* 80:2067–2076
  37. Simoben CV, Qaseem A, Moumbock AFA, Telukunta KK, Günther S, Sippl W, Ntie-Kang F (2020) Pharmacoinformatic investigation of medicinal plants from East Africa. *Mol Inf* 39:e2000163
  38. Stork C, Embruch G, Šícho M, de Bruyn KC, Chen Y, Svozil D, Kirchmair J (2020) NERDD: a web portal providing access to in silico tools for drug discovery. *Bioinformatics* 36:1291–1292
  39. Šícho M, Stork C, Mazzolari A, de Bruyn KC, Pedretti A, Testa B, Vistoli G, Svozil D, Kirchmair J (2019) FAME 3: predicting the sites of metabolism in synthetic compounds and natural products for phase 1 and phase 2 metabolic enzymes. *J Chem Inf Model* 59:3400–3412
  40. de Bruyn KC, Stork C, Šícho M, Kochev N, Svozil D, Jeliazkova N, Kirchmair J (2019) GLORY: generator of the structures of likely cytochrome P450 metabolites based on predicted sites of metabolism. *Front Chem* 7:402
  41. de Bruyn KC, Šícho M, Mazzolari A, Kirchmair J (2021) GLORYx: prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. *Chem Res Toxicol* 34:286–299
  42. Stork C, Chen Y, Šícho M, Kirchmair J (2019) Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. *J Chem Inf Model* 59:1030–1043
  43. Chen Y, Stork C, Hirte S, Kirchmair J (2019) NP-Scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomolecules* 9:43
  44. Wilm A, Norinder U, Agea MI, de Bruyn KC, Stork C, Kühnl J, Kirchmair J (2021) Skin Doctor CP: conformal prediction of the skin sensitization potential of small organic molecules. *Chem Res Toxicol* 34:330–344
  45. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G, Evelo CT, Guha R, Steinbeck C (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 9:33
  46. Sorokina M, Steinbeck C (2019) NaPLoS: a natural products likeness scorer-web application and database. *J Cheminform* 11:55
  47. Schaub J, Zielesny A, Steinbeck C, Sorokina M (2020) Too sweet: cheminformatics for deglycosylation in natural products. *J Cheminform* 12:67
  48. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Anal Chem* 78:779–787
  49. Tautenhahn R, Boettcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform* 9:504
  50. Benton HP, Want EJ, Ebbels TMD (2010) Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data. *Bioinformatics* 26:2488
  51. Verhoeven S, Schorn M, Willighagen E, van der Hooft J (2021) Paired omics data platform (version v0.9.2). Zenodo. <https://doi.org/10.5281/zenodo.4575489>
  52. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH (2017) plantSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res* 45:W55–W63
  53. Weber T, Kim HU (2016) The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production. *Synth Syst Biotechnol* 1:69–79
  54. Diallo BN, Glenister M, Musyoka TM, Lobb K, Tastan Bishop Ö (2021) SANCDB: an update on South African natural compounds and their readily available analogs. *J Cheminform* 13:37
  55. Ridder L, van der Hooft JJ, Verhoeven S, de Vos RC, van Schaik R, Vervoort J (2012) Substructure-based annotation of high-resolution multistage MS(n) spectral trees. *Rapid Commun Mass Spectrom* 26:2461–2471
  56. Ridder L, van der Hooft JJ, Verhoeven S (2014) Automatic compound annotation from mass spectrometry data using MAGMa. *Mass Spectrom* 3:50033
  57. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res* 48:W395–W402
  58. Moumbock AFA, Gao M, Qaseem A, Li J, Kirchner PA, Ndingkokhar B, Bekono BD, Simoben CV, Babiaka SB, Malange YI, Sauter F, Zierep P, Ntie-Kang F, Günther S (2021) StreptomeDB 3.0: an updated compendium of streptomycetes natural products. *Nucleic Acids Res* 49:D600–D604
  59. Valli M, dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2013) Development of a natural products database from the biodiversity of Brazil. *J Nat Prod* 76:439–444

60. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2017) NuBBE<sub>DB</sub>: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep* 7:7215
61. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, Juan-Blanco T, Aloy P (2020) Extending the small-molecule similarity principle to all levels of biology with the chemical checker. *Nat Biotechnol* 38:1087–1096
62. Sánchez-Cruz N, Medina-Franco JL (2021) Epigenetic target profiler: a web server to predict epigenetic targets of small molecules. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.1c00045>
63. Sánchez-Cruz N, Pílon-Jiménez BA, Medina-Franco JL (2020) Functional group and diversity analysis of BIOFACQUIM: a Mexican natural product database. *F1000Research* 8:2071. <https://doi.org/10.12688/f1000research.21540.2>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

