

Computational Metabolomics: Towards Molecules, Models, and their Meaning

Timothy M. D. Ebbels^{*1}, Soha Hassoun^{*2}, Ewy A. Mathé^{*3},
Justin J. J. van der Hooft^{*4}, and Haley Chatelaine^{†5}

1 Imperial College London, GB. t.ebbels@imperial.ac.uk

2 Tufts University – Medford, US. soha@cs.tufts.edu

3 National Institutes of Health – Bethesda, US. ewy.mathe@nih.gov

4 Wageningen University and Research, NL. justin.vanderhooft@wur.nl

5 NCATS – Bethesda, US. haley.chatelaine@nih.gov

Abstract

Dagstuhl Seminar “Computational Metabolomics: Towards Molecules, Models, and their Meaning,” (24181) is the fifth edition of the Computational Metabolomics seminars. Experts in fields ranging from cheminformatics, computer science, bioinformatics, analytical chemistry, and epidemiology attended to address the current state and future directions of this multi-disciplinary field. Specific topics of discussion were decided by participants but largely revolved around the seminar’s titular themes of molecules (i.e., utilizing and annotating individual metabolites for use in models), models (i.e., generating systems from which to derive meaning), and meaning (i.e., deriving actionable insights to further understanding of biological systems). New to this seminar, topics of education and training, as well as the use of large language models to enhance access to resources, were also discussed. Participants identified community needs for a balance of standardization and flexibility in realms of repository-scale data deposition and analysis, spectral library generation, automation best practices, and biological pathway interpretation. Participants also identified a number of action items toward these ends, fostering international collaborations among them. For example, one topic evolved around creating a benchmarking dataset for structure annotation based on MS/MS spectral data. Discussions represented balanced perspectives, thanks to varied session facilitators and active participation of all members. The report contained herein reflects highlights of each session, including informal evening sessions and ideas for future directions.

Seminar April 28 – May 3, 2024 – <https://www.dagstuhl.de/24181>

2012 ACM Subject Classification Applied computing → Life and medical sciences

Keywords and phrases bioinformatics, cheminformatics, data integration, machine learning, mass spectrometry, metabolite identification, metabolomics, pathway analysis, repository-scale analysis, training and education

Digital Object Identifier 10.4230/DagRep.14.4.124

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Computational Metabolomics: Towards Molecules, Models, and their Meaning, *Dagstuhl Reports*, Vol. 14, Issue 4, pp. 124–141

Editors: Timothy M. D. Ebbels, Soha Hassoun, Ewy A. Mathé, and Justin J. J. van der Hooft



DAGSTUHL Dagstuhl Reports

REPORTS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany


1 Executive Summary

Timothy M. D. Ebbels (Imperial College London, GB)

Soha Hassoun (Tufts University – Medford, US)

Ewy A. Mathé (National Institutes of Health – Bethesda, US)

Justin J. J. van der Hooft (Wageningen University and Research, NL)

License  Creative Commons BY 4.0 International license
© Timothy M. D. Ebbels, Soha Hassoun, Ewy A. Mathé, and Justin J. J. van der Hooft

Metabolomics is the study of the small molecule composition of biological systems. These small molecules define biological functions, making metabolomics a broadly applied technology in the biomedical, environmental, and biotechnology fields of study. Metabolite measurements are typically produced using mass spectrometry (MS), usually coupled with liquid or gas chromatography (LC or GC), and/or nuclear magnetic resonance (NMR) spectroscopy. New technologies continue to be developed, leading to increased resolution of metabolite species detected, as well as increased sensitivity. The resulting datasets are high throughput and typically yield abundances of thousands of small chemical structures. For these reasons, the field of computational metabolomics continues to grow to address current and imminent issues in data stewardship, processing, analysis and interpretation.

This Dagstuhl Seminar is the 5th seminar in the computational metabolomics series. Previous seminars have addressed key topics in the field. These include leveraging spectral data to annotate and identify routinely measured metabolites, assessing interactions between metabolites and proteins through “metaboproteomic assays”, as well as implementing multi-omic analyses and interpreting these data through enrichment analyses. This year, we not only dug deeper into the most current and relevant topics but also introduced new topics to the series. Our overall goal was to explore how to improve the utility of metabolomics data and its scientific relevance across many disciplines, alone or in combination with other data types, by leveraging machine learning and deep learning (ML/DL). To accomplish this, our seminar was organized into four categories. The first category was education, which was a newly introduced topic for this edition. Participants recognized the need for resources linking out to available education and training materials and discussed the inherent challenges of teaching multi-disciplinary topics, such as computational metabolomics. Our second category was molecules, which includes annotations and measurements of metabolites and molecules they associate with, such as proteins, genes, and other metabolites. New areas of emphasis included representation and classification of lipids, polymers and multi-constituent substances, and use of electron-activated disassociation methods for data collection. Our third category was models, which encompasses data quality, uncertainty in annotating metabolites, and relationships between metabolites and other molecules. Sessions in this category were the most prominent and included novel areas of repository-scale analyses, simulations of metabolomic data, and automation of data analysis workflows. Of note, practical sessions on how to best measure scientific impact and how/where to submit data publicly to increase utility of data and ability to develop new computational methods were held. Lastly, our fourth category was meaning, which represents resources, methods and tools that enable visualization and interpretation of large-scale metabolomic data in the context of biological, environmental or other sciences. This included the use and misuse of molecular networking and its current applications in the field. This year, new focus was placed on building interpretable models and leveraging increasingly available information on metabolite annotations, such as pathways and reactions. Of note, recent developments in

large language model techniques were discussed throughout our categories, with a special session dedicated to prototyping an LLM with metabolomics-specific content that can be used for training the next generation of metabolomics experts.

As in previous years, and based on positive feedback, the seminar format was again flexible, and topics were finalized on the first day of the seminar. The audience was encouraged to bring forth topics, and attention was paid to rotate who moderated and took notes for the sessions. Moderators, as well as organizers, actively ensured that all participants had a voice in topic sessions. Due to the large number of topics and to keep the groups manageable in size, parallel sessions were held. At the end of each day, the last group meeting was held to summarize the day's discussion and to finalize the planning for the following day. One new aspect this year was the set-up of a Slack workspace that was used prior to, during, and after the seminar. This workspace facilitated communication and information exchange between participants. Overall, this seminar was highly successful, and participants were highly engaged. Topics and discussions generated much enthusiasm and concrete next steps for potential collaborations. The field of computational metabolomics is a very active field of study that is somewhat underrepresented in many of the main metabolomics conferences, especially when it comes to tool development. The opportunity to focus on the computational aspects was very well received and will surely continue to grow as the data generation and types are ever increasing.

2 Table of Contents

Executive Summary

Timothy M. D. Ebbels, Soha Hassoun, Ewy A. Mathé, and Justin J. J. van der Hooft . 125

Overview of Talks

Biomarker discovery: current status of the field and experienced limitations <i>Carl Brunius and Purva Kulkarni</i>	129
Generative molecular models <i>Roman Bushuiev and David Wishart</i>	129
Reporting knowledge and accounting for uncertainty in machine learning for metabolite annotations <i>Roman Bushuiev and Sebastian Böcker</i>	130
Degree of automation – misuse of tools, degrees of uncertainty, traceability, appropriate descriptors, etc. <i>Haley Chatelaine and Ewy A. Mathé</i>	130
Visualization and networks for improved interpretability <i>Ronan Daly and Timothy M. D. Ebbels</i>	131
Combining unsupervised and supervised for metabolite annotation <i>Niek de Jonge and Justin J. J. van der Hooft</i>	131
Education in computational metabolomics <i>Timothy M. D. Ebbels, Ewy A. Mathé, Stacey N. Reinke, and Denise Slenter</i>	132
Polymers and multi-constituent substances <i>Tytus Mak, Emma Schymanski, and Egon Willighagen</i>	132
Building interpretable models using pathways <i>Ewy A. Mathé and Timothy M. D. Ebbels</i>	133
Simulation of metabolomics data <i>Ewy A. Mathé, Ronan Daly, and Stacey N. Reinke</i>	133
Using pathways/networks/reactions for metabolite identification <i>Hosein Mohimani</i>	134
Computational methods for biomarker discovery <i>María Eugenia Monge, Daniel Raftery, and Denise Slenter</i>	134
Molecular networking in the GNPS environment: applications, considerations, and future directions <i>Raphael Reher and Justin J. J. van der Hooft</i>	135
LLMs <i>Stacey N. Reinke and Nicola Zamboni</i>	135
Multi-omics integration <i>Stacey N. Reinke and Juan Antonio Vizcaino</i>	136
Mass spectral reference library generation and mass spectral quality <i>Robin Schmid and Tytus Mak</i>	136
Using networks/reactions for metabolic networks and data interpretation <i>Denise Slenter, Timothy M. D. Ebbels, and Egon Willighagen</i>	137

Towards MS/MS annotation benchmark at NeurIPS 2024 <i>Michael Andrej Stravs and Roman Bushuiev</i>	137
Access to public data and how it should be submitted <i>Juan Antonio Vizcaino, Alice Limonciel, and Ewy A. Mathé</i>	138
Big data in metabolomics: repository-scale analyses <i>Juan Antonio Vizcaino and Ralf Weber</i>	138
Scientific impact: Methods for article citations and annotations of why you're being cited <i>Egon Willighagen and Carl Brunius</i>	139
Structural representation of lipid classes and enumeration <i>Egon Willighagen and Michael Anton Witting</i>	140
Beyond Collision-Induced Dissasication with Electron-Activated Disassociation <i>Nicola Zamboni</i>	140
Participants	141

3 Overview of Talks

3.1 Biomarker discovery: current status of the field and experienced limitations

Carl Brunius (Chalmers University of Technology – Göteborg, SE) and Purva Kulkarni (Radboud University – Nijmegen, NL)

License © Creative Commons BY 4.0 International license
© Carl Brunius and Purva Kulkarni

The session on “Combining Unsupervised and Supervised Models for biomarker discovery” was renamed to “Biomarker discovery: current status of the field and experienced limitations” after redefining the scope. The session started with defining the meaning of a biomarker, which is not a homogenous concept and brought forward different perspectives from the attendees.

The fact that the importance of biomarker validation should be significantly more than that of biomarker discovery was brought up since many biomarkers do not get approved easily. The scope for biomarker validation-related discussion included experimental and computational validation, application of analytical assays to determine the stability, reproducibility, effectiveness and possible costs for clinical translation. It also came up during the discussion that a lot is reported in repositories about biomarkers related to central metabolism but not much about metabolic biomarkers for exposure (exposomics studies). It was identified that there is a need for tools and databases to study drug mechanisms, possible side effects and consequences of changes in metabolism.

Regarding using metabolomics for biomarker discovery, several databases were discussed that contain reference values for listed metabolites. This was mainly in the context of metabolites commonly detected in bio-fluids over the life course. Apart from this, an initiative like the Consortium of Metabolomics Studies (COMETS) was brought up that contains information on available population-based cohorts. Additionally, it would be beneficial to know about publicly available datasets that have been used in the past for biomarker validation. It was discussed that, overall, biomarker research and translation into clinical practice needs to focus on increasing simplicity and stability and hence could focus more on targeted assays rather than untargeted metabolomics approaches.

In conclusion, the session emphasized the importance of biomarker validation and identifying approaches to have stable effective biomarkers, not just for disease diagnosis and therapy monitoring but also for risk prediction and prevention. Additionally, the discussion stressed the importance of exposure biomarkers from a more general epidemiological perspective, since exposure assessment is key to achieving relevant exposure-outcome risk assessments.

3.2 Generative molecular models

Roman Bushuiev (The Czech Academy of Sciences – Prague, CZ) and David Wishart (University of Alberta – Edmonton, CA)


License © Creative Commons BY 4.0 International license
© Roman Bushuiev and David Wishart

In this session, we focused on the advancements and challenges in generating molecular structures from MS/MS spectra. We discussed various generative methods, including chemical language models based on SMILES strings and reinforcement learning approaches such as

the Reinvent model. The discussion highlighted the need for new, more relevant metrics that reflect the ambiguity of a predicted molecular structure and new conditioning mechanisms that effectively leverage mass spectra. We explored related work in the field of drug design and discussed the relevance of mass spectrometry data beyond MS/MS. Insights from the session underscored the importance of de novo generation of molecules from mass spectra and identified key challenges.

3.3 Reporting knowledge and accounting for uncertainty in machine learning for metabolite annotations


Roman Bushuiev (The Czech Academy of Sciences – Prague, CZ) and Sebastian Böcker (Friedrich-Schiller-Universität Jena, DE)

License  Creative Commons BY 4.0 International license
© Roman Bushuiev and Sebastian Böcker

This session delved into the challenge of reporting uncertainty in machine learning-based metabolite annotations from mass spectra. Discussions highlighted the importance of users understanding prediction uncertainty, particularly when dealing with spectra beyond the training data distribution. The focus was on the SIRIUS software, which estimates uncertainty by explaining spectral peaks based on a predicted molecular fingerprint. We explored user-friendly options for handling reported uncertainty, such as setting thresholds and enabling user-controlled visualization. Two main use cases were identified: identifying single compounds and conducting large-scale downstream statistical analysis, each requiring tailored uncertainty management. We brainstormed potential new methods for estimating uncertainty, including perturbation analysis of predicted structures and in silico fragmentation. Finally, we discussed uncertainty in the context of the undiscovered chemical space.

3.4 Degree of automation – misuse of tools, degrees of uncertainty, traceability, appropriate descriptors, etc.

Haley Chatelaine (NCATS – Bethesda, US) and Ewy A. Mathé (National Institutes of Health – Bethesda, US)

License  Creative Commons BY 4.0 International license
© Haley Chatelaine and Ewy A. Mathé

Automation holds strong promise in minimizing inevitable errors in collecting sample meta-data, protocol meta-data, and data processing parameters. This session involved a discussion of participants' existing use of automation, hurdles, and insights into necessities for best practices for its use on the aforementioned levels of study design. The traceability of automation, particularly when used in tandem with lab information management systems (LIMS), facilitates reproducibility. However, a clear core of common concepts that can be conserved across all metabolomics labs needs to be delineated so that labs know how they can optimize protocols for their unique fit-for-purposes. A repository of existing tools and best practices could also help facilitate this process. This is all with the caveat that automated processes need to incorporate appropriate diagnostics to identify sources of error in the process.

3.5 Visualization and networks for improved interpretability

Ronan Daly (University of Glasgow – Bearsden, GB) and Timothy M. D. Ebbels (Imperial College London, GB)

License © Creative Commons BY 4.0 International license
© Ronan Daly and Timothy M. D. Ebbels

This session mainly focused on the direct visualization of networks related to metabolomics (and possibly multi-omics). These networks tend to have metabolites as nodes, with edges showing a relationship between them. Two of the main sources of network topology are knowledge driven, from bases such as KEGG or Reactome, and data driven using edge sources such as correlation or differential-correlation analyses. The “hairball” effect is known to be endemic in these data-based analyses, with methods such as partial correlation and filtering used to mitigate this. Tools to display metabolic networks were noted to be in short supply; programs such as Cytoscape and neo4j are used but do not cover all needs. New, easy-to-use tools that can flexibly display a range of different networks are needed. These should be able to use diverse knowledge bases as input, whilst being flexible as to other data that can be displayed, including annotating nodes and edges with complex data types and having layouts that are robust to initialisation. Including knowledge- and data-driven nodes and edges would be useful to display well-known and novel connections.

3.6 Combining unsupervised and supervised for metabolite annotation

Niek de Jonge (Wageningen University, NL) and Justin J. J. van der Hooft (Wageningen University and Research, NL)

License © Creative Commons BY 4.0 International license
© Niek de Jonge and Justin J. J. van der Hooft

This session discussed how we could combine unsupervised models with supervised models for annotating MS2 mass spectra and the potential of using multistage MS_n mass fragmentation spectra as an input for such models. So far there have been trained unsupervised models (e.g., Spec2Vec, MS2LDA, and now very recently DreaMS) and supervised models such as MS2DeepScore. However, the approach of fine tuning unsupervised models has not been common for mass spectrum annotation but could be very promising. We discussed examples in other fields, like image labeling, and discussed how we can apply these principles to mass spectrum annotation. We have two concrete plans on how to do this. One approach would be to link mass spectral embeddings with molecular embeddings, the other approach is to fine tune unsupervised models to perform specific tasks like predicting chemical similarity. During the discussion, we highlighted potential routes, opportunities, and pitfalls when doing this. Overall, the group feels that this is a very promising avenue for further research where there will be place for both unsupervised and supervised models and combinations thereof.

3.7 Education in computational metabolomics

Timothy M. D. Ebbels (Imperial College London, GB), Ewy A. Mathé (National Institutes of Health – Bethesda, US), Stacey N. Reinke (Edith Cowan University – Joondalup, AU), and Denise Slenter (Maastricht University, NL)

License © Creative Commons BY 4.0 International license
© Timothy M. D. Ebbels, Ewy A. Mathé, Stacey N. Reinke, and Denise Slenter

The aim of this session was to identify ways in which the community can provide education and training content to a varied audience. Education deals with long-term and broad content that focuses on concept development while training deals with short-term and focused content that is skills-focused. Education pieces are thus relatively static and do not change much over time while training pieces are constantly evolving. Challenges in metabolomic education and training include dealing with many types of users with different levels of education on metabolomics analyses, difficulties in producing documentation, and variability in awareness of available content and workshops (e.g., spread across the internet and countries and not always findable). Notably, the desire to bring the community together on providing consistent and reliable education and training content was strong. All participants saw the need to have a better overview of existing teaching materials available. Training people from different backgrounds (medical, chemistry, biology, computational, or a mixture) can be quite challenging. More embedding in existing curricula would be a good way to spark interest and enthusiasm for (computational) metabolomics. Having an online collection of existing materials, such as the Software Data Exchange through the Metabolomics Association of North America (MANA SODA), which is also checked for availability (e.g. R-package availability, function documentation) would help in keeping materials in line with the fast developments of software. Action items from this session include uniting the Metabolomics Society and Affiliates to consolidate available resources and piloting a chatbot using content that is provided from participants in this meeting.

3.8 Polymers and multi-constituent substances

Tytus Mak (NIST – Gaithersburg, US), Emma Schymanski (University of Luxembourg, LU), and Egon Willighagen (Maastricht University, NL)

License © Creative Commons BY 4.0 International license
© Tytus Mak, Emma Schymanski, and Egon Willighagen

A concerted effort is currently underway to create the computational infrastructure for supporting spectral data for multi-constituent substances (MCS), with the initial focus on polymers. NIST is in the early stages of acquiring data for building spectral libraries, with an initial focus on pyrolysis GC-MS but expanding to other instrument platforms such as LC-MS. NIST is also developing new algorithms to support the matching of polymers and polymer mixtures, which by their nature consists of a multitude of spectra at multiple retention times even for “pure” samples, rather than the one-compound-one-spectrum paradigm of traditional spectral libraries. Critical to this effort is the cheminformatics necessary to precisely define the molecular structures of polymers, and support via PubChem and other molecular databases. This effort is being led by collaborators at Maastricht University and University of Luxembourg.

3.9 Building interpretable models using pathways

Ewy A. Mathé (National Institutes of Health – Bethesda, US) and Timothy M. D. Ebbels (Imperial College London, GB)

License © Creative Commons BY 4.0 International license
© Ewy A. Mathé and Timothy M. D. Ebbels

Metabolomic data are difficult to interpret in terms of a biological “story.” Biological pathways, here defined as lists of molecules participating in a common function, can be useful for this. Pathway enrichment is one example, but there are many other ways of employing this information. This discussion focused on highlighting gaps and needs for improving pathway enrichment and interpretation of metabolomic data and for encouraging novel approaches to the problem. Examples include adding pathophysiological pathways (e.g. necrosis) to databases, improving completeness of metabolite-pathway mappings, propagating uncertainty (e.g. from annotation) through to the pathway level, and quantification of specificity of the pathway signal. Suggested action items included a review paper highlighting these gaps, benchmarking study(ies) assessing uncertainty in the pathway methods, and methodological work addressing the question of how to quantify specificity.

3.10 Simulation of metabolomics data

Ewy A. Mathé (National Institutes of Health – Bethesda, US), Ronan Daly (University of Glasgow – Bearsden, GB), and Stacey N. Reinke (Edith Cowan University – Joondalup, AU)

License © Creative Commons BY 4.0 International license
© Ewy A. Mathé, Ronan Daly, and Stacey N. Reinke

The aim of this session was to explore and define the uses of simulated data and how generating gold-standard experimental metabolomics data with ground truth could help in numerous tasks. Globally, simulated metabolomic data provide a ground truth against which concepts can be demonstrated and are thus useful for benchmarking. Different contexts and types of analyses require different simulated datasets, with the axis from raw to processed data being particularly important. Coupled to this is the fact that mass spectrometry data is very heterogeneous, with possibly very flexible tools needed. Currently, most researchers simulate data in an ad-hoc manner, thereby making comparison of tasks and their utility difficult. Many uses of simulated datasets were defined, including evaluating algorithmic pipelines, statistical inference, and teaching/education. While some publications describe simulation tools, they are relatively few, and there is a lack of a centralized repository for benchmark and simulated data. A review of approaches to simulating metabolomic data as well as an associated repository of simulation tools were suggested as useful action items.

3.11 Using pathways/networks/reactions for metabolite identification


Hosein Mohimani (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Hosein Mohimani

This session covered several angles on how the community can use pathways knowledgebases to aid in metabolite annotation. Structure-based approaches predict the product of enzymatic reactions solely based on the substrate structure and knowledge about the reaction host/environment and enzymes involved. Metabolomics-based approaches exist that group mass features and filter for known mass differences that correspond to biotransformations. The consensus was that it might be too early to explore whole reactome databases, and looking into single reactions should be considered instead, as a low hanging fruit. Additionally, training datasets are currently really small, and additional experimental data collection efforts are needed to improve them. These experimental data should be collected on simple systems (e.g., individual enzymes), but more complex systems (e.g., whole microbiome) would not be useful, as it would be very difficult to annotate them.

3.12 Computational methods for biomarker discovery


María Eugenia Monge (CIBION – Buenos Aires, AR), Daniel Raftery (University of Washington – Seattle, US), and Denise Slenter (Maastricht University, NL)

License  Creative Commons BY 4.0 International license
© María Eugenia Monge, Daniel Raftery, and Denise Slenter

This session focused on the accumulation of biomarker data, largely for use in clinical assays, and some of the challenges in this area, both in the sparsity of reliable biomarker information in existing databases and in the computational aspects in developing strong biomarker panels. There was a suggestion to focus more on the collection of large targeted assays that provide more reliable data and in particular, absolute concentrations. But some also recognized that untargeted metabolomics can deliver novel metabolite biomarkers. Reference ranges do exist for a large number of metabolites, and they are collected in MarkerDB. However, there are issues in these data, such as different sets of ranges that depend on experimental parameters, such that the ranges can be rather large. Regarding computational approaches, some suggestions were made for FDR correction and how metabolite classes (triglycerides, acylcarnitines) can be used instead of the individual metabolites. Similarly, ratios of metabolites are increasingly of interest, but methods to focus on the most important ratios are very important to avoid false discovery rate issues with the potentially huge number of potential ratios that could be calculated. It was agreed that more data need to be made available, possibly by soliciting researchers who have access to them, such as those who have already deposited data to the existing databases. Issues of confidentiality and patient consent may restrict these data to summaries, i.e., without individual-level metabolite values. Nevertheless, there seems to be some specific ways forward to improve the biomarker discovery and validation process, and better data quality and coverage will provide much better inputs for a variety of computational methods to develop strong biomarker panels.

3.13 Molecular networking in the GNPS environment: applications, considerations, and future directions


Raphael Reher (Universität Marburg, DE) and Justin J. J. van der Hooft (Wageningen University and Research, NL)

License  Creative Commons BY 4.0 International license
© Raphael Reher and Justin J. J. van der Hooft

Molecular networking is a powerful method for visualizing and annotating the chemical space in non-targeted mass spectrometry (MS) data. We noted that “mass fragmentation-based spectral grouping” was a more accurate name, but Molecular Networking is well established in the community. In this session, various aspects related to molecular networking, including applications, flavors, downstream tasks, and visualization tools, were discussed. The applications are wide: i) finding analogues, ii) creating a mindmap of the data, iii) mass feature annotation (using the network topology), and iv) finding small groups of highly-interconnected nodes representing xenobiotics. We concluded that for some applications the network as such is not necessarily needed. We concluded that Molecular Networking combines several tasks in one (organization and visualization) and has a broad variety of applications that each tap into either one or both of the tasks. We noted that the default settings are not suitable for most tasks; however, we do not have good ways yet of assessing what thresholds, settings, and similarity scores would work best for which scenarios. We discussed the use of different similarity metrics, including various mass spectral scores, but also the “shared substructures” (inferred by MS2LDA), fingerprints (inferred by SIRIUS), or biochemical distance. Furthermore, we discussed that graph-based approaches have been applied to mass spectral networks in a very limited fashion, and this could be an interesting route to explore as an alternative to the current spectral grouping.

3.14 LLMs


Stacey N. Reinke (Edith Cowan University – Joondalup, AU) and Nicola Zamboni (ETH Zürich, CH)

License  Creative Commons BY 4.0 International license
© Stacey N. Reinke and Nicola Zamboni

Following discussions in the Education and Training sessions earlier in the week, Dagstuhl attendees in this session provided content, and an LLM prototype was designed. In this session, a demonstration of the prototype was conducted. We discussed challenges, considerations, and what is needed to improve the prototype in the future.

3.15 Multi-omics integration

Stacey N. Reinke (Edith Cowan University – Joondalup, AU) and Juan Antonio Vizcaino (EMBL-EBI – Hinxton, GB)

License  Creative Commons BY 4.0 International license
© Stacey N. Reinke and Juan Antonio Vizcaino

Multi-omics integration refers to analyzing more than one set of omics data to create a more comprehensive understanding of a biological system. Use cases of multi-omics integration vary from functional biochemistry in model systems to discovery-driven analysis in human studies. Several approaches, methods, and tools exist to integrate omics data; however, these are often complex and can be inaccessible to non-experts. Other challenges of performing multi-omics integration include a high data dimensionality (many variables), inherent bias to higher variable data blocks, and complexity of results. Future directions include making methods and tools more accessible so that a wider range of researchers can use them.

3.16 Mass spectral reference library generation and mass spectral quality

Robin Schmid (The Czech Academy of Sciences – Prague, CZ) and Tytus Mak (NIST – Gaithersburg, US)

License  Creative Commons BY 4.0 International license
© Robin Schmid and Tytus Mak


This session recognized and discussed current issues concerning the generation of mass spectral reference libraries in the metabolomics “small” molecule space. Current libraries are static and often lack a reference to the original raw data. Furthermore, data (spectral) processing is done without providing reproducible processing parameters.

Briefly, the library generation workflow should be streamlined and made accessible to experimental mass spectrometrist, by abstracting many steps, including (a) compound and experimental metadata curation by structure cleanup and database lookup, (b) providing recommendations and SOPs for popular MS platforms and methods to guide data acquisition, and (c) create a Github repository to collect resources and tools for data processing and spectral quality scoring for library generation

Finally, we are planning to connect these tools into a workflow and plan the infrastructure and possible front ends. The remaining key issues are defining the required and optional metadata to be captured. We agreed that automatic processing and library generation has limitations, and manual curation needs to be guided and tracked, to preserve expert knowledge. There was a consensus that the 8 action items discussed during our session would significantly impact the mass spectrometry community if implemented. The 8 action items included: (1) contributing tools for depositing meta-data in the library, (2) defining a minimum requirement for compound and experimental meta-data, (3) documenting best practices for MS data processing for library generation, (4) listing libraries and databases for data pushes, (5) documenting who is open to sharing compounds, (6) creating a layout of software design and infrastructure, (7) setting up an online meeting for interested parties, and (8) creating a repository of spectral libraries and collecting feedback to verify interest in sending standard aliquots for library generation. We could easily increase the public MS libraries n-fold by inviting analytical chemists.

3.17 Using networks/reactions for metabolic networks and data interpretation

Denise Slenter (Maastricht University, NL), Timothy M. D. Ebbels (Imperial College London, GB), and Egon Willighagen (Maastricht University, NL)

License  Creative Commons BY 4.0 International license
© Denise Slenter, Timothy M. D. Ebbels, and Egon Willighagen

Chemists appreciate the individual biochemical reactions covered in various databases. Combining all of these reactions in one network can be useful to overcome the relatively arbitrary boundaries inherent in categorizing biochemical pathways. However, the categorisation of reactions in pathways can be very relevant for the biological interpretation of (metabol)omics data.

Several issues were brought up that influence the usability of the currently captured machine-readable reaction knowledge. To name a few: the utility of capturing (missing/unknown) stereochemistry, mutable identification annotations, the ability to merge data based on name matching, the frequent inability to reuse existing models, low metabolite coverage, unknown metabolite biotransformations, and unconnected reactions.

Furthermore, the tools and techniques developed for other -omics fields are not directly applicable for metabolomics, due to sparsity and coverage issues. Finding reactions that are not in equilibrium from metabolomics data and changes thereof is still complicated today, and teaching materials and integration thereof in relevant curricula is lacking.

3.18 Towards MS/MS annotation benchmark at NeurIPS 2024


Michael Andrej Stravs (Eawag – Dübendorf, CH) and Roman Bushuiev (The Czech Academy of Sciences – Prague, CZ)

License  Creative Commons BY 4.0 International license
© Michael Andrej Stravs and Roman Bushuiev

In this session, we discussed questions associated with the goal to submit a dataset and benchmark for computational metabolomics to NeurIPS 2024. The principal motivation is to attract interest from the machine learning community. For this purpose, the dataset and tasks must pose a minimal barrier of entry. It was decided to focus on a dataset with columns of molecular structure, m/z, intensities and collision energy, with detailed datatype descriptions. Tasks should be formulated as prediction of specific columns as labels from a set of input columns. The tasks should include structure-to-spectrum prediction and spectrum to candidate list ranking. Disagreement reigned about whether a surrogate task such as fingerprint prediction should be included. Further, de novo structure generation was discussed as an attractive, but hard-to-evaluate, task. As another issue, the test-train split needs to be considered carefully to avoid “too similar” molecules in separate folds; MCES-based or scaffold-based splits were discussed. Importantly, naive baselines such as k-nearest neighbor are required to contextualize model performance.

3.19 Access to public data and how it should be submitted

Juan Antonio Vizcaino (EMBL-EBI – Hinxton, GB), Alice Limonciel (biocrates life sciences – Innsbruck, AT), and Ewy A. Mathé (National Institutes of Health – Bethesda, US)

License  Creative Commons BY 4.0 International license
© Juan Antonio Vizcaino, Alice Limonciel, and Ewy A. Mathé

In this session, we discussed the most frequent use cases of data reuse today: reproducing the result of a study, meta-analysis studies, reanalysing untargeted datasets to find new compounds, reanalysis with a different objective than the one from the original study, and machine-learning approaches using the mass spectra as the basis. Apart from GNPS-related efforts, there are limited examples about data reuse in metabolomics

If we had to choose among different types of metadata, it was mentioned that biological information was preferable since the methodological information is more difficult to capture in high-detail and most often it is required to read the manuscript.

The possibility of having a two-tier system for data submissions was discussed to increase the amount of data in the public domain. The people doing the higher-tier submission should get an incentive: a DOI, extra curation and extra promotion (“dataset of the week”, or something similar).

It was also felt afterwards in the plenary session that we still should aspire to publish and share data in a way that enables cross-dataset (meta)studies and repository-scale analyses

3.20 Big data in metabolomics: repository-scale analyses

Juan Antonio Vizcaino (EMBL-EBI – Hinxton, GB) and Ralf Weber (University of Birmingham, GB)


License  Creative Commons BY 4.0 International license
© Juan Antonio Vizcaino and Ralf Weber

The session on “Big data in metabolomics: repository-scale analyses” was well-attended, with people coming from multiple backgrounds. Discussions started with some current examples of data reuse, ranging from re-processing raw data, annotation and discovery of metabolites, machine-learning approaches to train models and the (lack of) meta-studies in metabolomics. These efforts surfaced some issues, both on a technical level and around (missing) metadata annotation and interruptions in the chain of evidence linking from raw data to annotated metabolites. In proteomics, the evidence chain goes backwards from the identified proteins and peptides, to the MS/MS spectra used in the identification and via the universal spectrum identifier (USI) directly to a spectrum in a deposited raw data file.

A possible carrot to improve the situation can (as always) be improved tooling, allowing to improve capturing metadata earlier in the process and export to the repositories, and, as an even bigger carrot, analysis software like MetaboAnalyst.

3.21 Scientific impact: Methods for article citations and annotations of why you're being cited

Egon Willighagen (Maastricht University, NL) and Carl Brunius (Chalmers University of Technology – Göteborg, SE)

License  Creative Commons BY 4.0 International license
© Egon Willighagen and Carl Brunius

In this session, Egon Willighagen introduced the topic with the history of the Citations Typing Ontology (CiTO), from the publication in 2010 [1], via the Journal of Cheminformatics pilot [2] and BioHackrXiv [3], to the indexing of the annotations in Wikidata and visualized by Scholia (link 1). The CiTO allows annotating citations with the intention of the citation. For example, you can indicate that you cite an article because you use data or database described in that article or that you use a method introduced in that article. It also allows you to indicate that you agree or disagree with that article, or that you ridicule that article. Using the annotations to create citation networks that reflect research reuse was discussed. For example, we can track the history of a software library (like the Chemistry Development Kit) or a database (like RaMP-DB) by following “cito:extends” citations. And when thinking about the impact of an article, citations by articles that reuse your work show more impact than citations that cite you as “related work.” During the meeting, we looked at how nanopublications can allow authors to provide citation intention annotations after publication (where J. Cheminform. and BioHackrXiv allowed adding this annotation as part of the publication itself). Nanopublications are an approach using linked data [4] to publish a single fact (hence a “nano”-publication) but with full provenance. We looked at a recently developed template (link 2) that provides a graphical user interface. It requires the author of the nanopublication to authenticate with their ORCID account. This latter point is considered essential to the participants, who realize anyone can make nanopublications. The six participants wrote several nanopublications reflecting eight annotated citations (link 3). One of these gives the intent of one citation that it cites the article for information (cito:citesForInformation) (link 4).

Links:


1. scholia.toolforge.org/cito
2. <https://shorturl.at/x5MRB>
3. <https://shorturl.at/RPZHZ>
4. <https://shorturl.at/nutoC>

References

- 1 <https://doi.org/10.1186/2041-1480-1-S1-S6>
- 2 <https://doi.org/10.1186/s13321-023-00683-2>
- 3 <https://doi.org/10.5281/zenodo.10072013>
- 4 <https://doi.org/10.1109/escience.2018.00024>

3.22 Structural representation of lipid classes and enumeration


Egon Willighagen (Maastricht University, NL) and Michael Anton Witting (Helmholtz Zentrum München, DE)

License  Creative Commons BY 4.0 International license
© Egon Willighagen and Michael Anton Witting

Lipids cover a large combinatorial space based on different backbones, headgroups and fatty acyls. Current tools for the analysis of structural details in lipids only allow a limited depth in the annotation of structural features, such as location of functional groups, position and stereochemistry of double bonds, position of hydroxyl groups, etc. New fragmentation technologies such as EAD, OAD or UPVD allow to delve deeper into structural annotation. However, still some uncertainty remains. This session focused on ways to capture this uncertainty in chemical representations such as SMILES or InChIs. This is a problem not only specific to lipids, but also many other molecule classes. Extended SMILES and a new implementation for InChI isotopologue and isotopomer specifications have been discussed. Solutions will allow investigators, in the future, to potentially report structures together with uncertainty measures. Further action points include the collection of explicit real-life examples to further optimize the definitions that will be required for capturing uncertainty in structures.

3.23 Beyond Collision-Induced Dissociation with Electron-Activated Disassociation

Nicola Zamboni (ETH Zürich, CH)

License  Creative Commons BY 4.0 International license
© Nicola Zamboni

The discussion of this session pivoted on the opportunities and needs associated with the adoption of electron-induced dissociation (aka EAD or EIEIO) for structural annotation of molecules. Much work has been devoted to lipids, and the results are far beyond expectations. It was discussed how to best address the analysis of different classes of molecules by the use of computational and ML methods. This includes (i) sharing EID data of representative classes with computational groups, (ii) measuring the information content of such spectra and comparing to CID, (iii) evaluating how CID-centric ML tools deal with the specific characteristics of EID spectra.

Participants

- Wout Bittremieux
University of Antwerp, BE
- Sebastian Böcker
Friedrich-Schiller-Universität
Jena, DE
- Carl Brunius
Chalmers University of
Technology – Göteborg, SE
- Roman Bushuiev
The Czech Academy of Sciences –
Prague, CZ
- Haley Chatelaine
NCATS – Bethesda, US
- Ronan Daly
University of Glasgow –
Bearsden, GB
- Niek de Jonge
Wageningen University, NL
- Kai Dührkop
Friedrich-Schiller-Universität
Jena, DE
- Timothy M. D. Ebbels
Imperial College London, GB
- Soha Hassoun
Tufts University – Medford, US
- Florian Huber
Hochschule Düsseldorf, DE
- Pär Jonsson
Sartorius Stedim Data Analytics –
Umeå, SE
- Purva Kulkarni
Radboud University –
Nijmegen, NL
- Jessica Lasky-Su
Brigham and Women’s Hospital
& Harvard Medical School –
Boston, US
- Alice Limonciel
biocrates life sciences –
Innsbruck, AT
- Liping Liu
Tufts University – Medford, US
- Tytus Mak
NIST – Gaithersburg, US
- Ewy A. Mathé
National Institutes of Health –
Bethesda, US
- Hosein Mohimani
Carnegie Mellon University –
Pittsburgh, US
- María Eugenia Monge
CIBION – Buenos Aires, AR
- Steffen Neumann
IPB – Halle, DE
- Louis-Felix Nothias
CNRS & Université Côte d’Azur
– Nice, FR
- Daniel Raftery
University of Washington –
Seattle, US
- Raphael Reher
Universität Marburg, DE
- Stacey N. Reinke
Edith Cowan University –
Joondalup, AU
- Hannes Röst
University of Toronto, CA
- Juho Rousu
Aalto University, FI
- Robin Schmid
The Czech Academy of Sciences –
Prague, CZ
- Emma Schymanski
University of Luxembourg, LU
- Denise Slenter
Maastricht University, NL
- Jan Stanstrup
University of Copenhagen, DK
- Michael Andrej Stravs
Eawag – Dübendorf, CH
- Marynka
Ulaszewska-Tarantino
Thermo Fisher Scientific –
Milan, IT
- Justin J. J. van der Hoof
Wageningen University &
Research, NL
- Dries Verdegem
VIB – KU Leuven, BE
- Juan Antonio Vizcaino
EMBL-EBI – Hinxton, GB
- Ralf Weber
University of Birmingham, GB
- Egon Willighagen
Maastricht University, NL
- David Wishart
University of Alberta –
Edmonton, CA
- Michael Anton Witting
Helmholtz Zentrum
München, DE
- Nicola Zamboni
ETH Zürich, CH

