doi:10.2533/chimia.2024.525 Chimia 78 (2024) 525–530

A Decade of Computational Mass Spectrometry from Reference Spectra to Deep Learning

Michael A. Stravs*

METAS Award 2023

Abstract: Computational methods are playing an increasingly important role as a complement to conventional data evaluation methods in analytical chemistry, and particularly mass spectrometry. Computational mass spectrometry (CompMS) is the application of computational methods on mass spectrometry data. Herein, advances in CompMS for small molecule chemistry are discussed in the areas of spectral libraries, spectrum prediction, and tentative structure identification (annotation): Automatic spectrum curation is facilitating the expansion of openly available spectral libraries, a crucial resource both for compound annotation directly and as a resource for machine learning algorithms. Spectrum prediction and molecular fingerprint prediction have emerged as two key approaches to compound annotation. For both, multiple methods based on classical machine learning and deep learning have been developed. Driven by advances in deep learning-based generative chemistry, *de novo* structure generation from fragment spectra is emerging as a new field of research. This review highlights key publications in these fields, including our approaches RMassBank (automatic spectrum curation) and MSNovelist (*de novo* structure generation).

Keywords: Machine learning · Mass spectrometry · Small molecules

Michael Stravs obtained a master's degree in Biology from ETH Zürich in 2011 and conducted his doctorate at Eawag from 2013–2017, graduating with a PhD from ETH Zürich in 2017. He was awarded the METAS Award 2023 for his work in computational and instrumental high-resolution mass spectrometry. Before and throughout his PhD research at Eawag with Prof. Juliane Hollender and Prof. Francesco Pomati, he conducted work supporting the establishment of the MassBank spectral library with Dr. Emma Schymanski. As a postdoc at Eawag with Dr. Christoph Ort and Heinz Singer, he developed the MS2field system, a fully automated transportable high-resolution mass spectrometry lab-in-a-trailer for water analysis. As an associate senior scientist at ETH Zürich at the Zamboni lab, he developed MSNovelist, an algorithm for *de novo* structure assignment from MS2 spectra. He is now working as a staff data scientist at Eawag.

1. Introduction

High-resolution mass spectrometry (HRMS) has become a key technology in environmental, forensic and life sciences. With time-of-flight (TOF) and Orbitrap instruments masses of organic molecules can be determined at low ppm-level mass accuracy, which enables tentative assignment of molecular formulae to ions. Using electrospray ionization (ESI) as a 'soft' ionization technology, molecules are ionized with minimal fragmentation to measure the (*pseudo*-)molecular mass. Fragmentation (MS/MS) can be induced for selected ions in collision cells (high-energy collision-induced dissociation, HCD), generating fragments for substructure characterization. Broadly speaking, HRMS applications are concerned either with organic molecules of arbitrary structures with masses up to 1,000–1,500 Da ('small molecules', *e.g.* in metabolomics or environmental chemistry) or with peptides and proteins, which have a linear structure composed of recurring building blocks (proteomics) and an expected mass range of 500–3,000 Da.

The field of *computational mass spectrometry* (*CompMS*) is concerned with generating knowledge from mass spectrometric data. Herein, I specifically discuss CompMS related to small molecule chemistry, while noting the large research body in proteomics related CompMS. Arguably, the central challenge in small molecule CompMS is annotation (Fig. 1), *i.e.* assigning a chemical structure to a detected feature, characterized by a mass, its fragmentation spectrum, and potentially additional data such as chromatographic retention time, isotopic pattern, ion mobility data, *etc.* In contrast to traditional targeted approaches, where known analytes are detected and quantified, the goal of 'nontarget analysis' and annotation is the discovery of untargeted chemicals in a sample. It should be acknowledged that even if all processes are perfectly understood, some molecules will remain indistinguishable with mass spectrometry (trivially, stereoisomers, but also an important fraction of regioisomers). All CompMS approaches are constrained by this analytical challenge; advances over the last decade have enhanced our understanding of the capabilities and limits of these methods.

2. Spectral Library Search and MassBank

2.1 *Spectral Libraries for ESI-HRMS*

The most straightforward approach for compound identification from mass spectra is searching for matches in a spectral library, where a measured spectrum is compared with a library

*Correspondence: Dr. M. A. Stravs, E-mail: stravsmi@eawag.ch
Eawag, Ueberlandstrasse 133, CH-8600 Dübendorf.

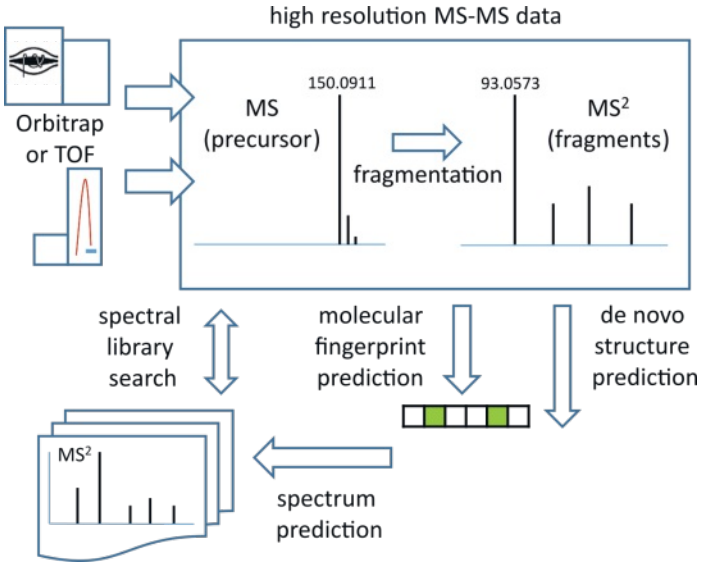

Fig. 1. Compound annotation approaches in CompMS. High-resolution MS/MS data is acquired on an Orbitrap or TOF instrument. Spectral library search: MS2 spectra are compared with a database of reference spectra. Spectrum prediction: Reference spectra databases are augmented with simulated spectra. Molecular fingerprint prediction: Substructural features are predicted from the MS2 spectrum for database search. *De novo* structure prediction: A chemical structure is generated according to the information in the MS2 spectrum.

of reference spectra of known analytes. For gas chromatography - electron impact mass spectrometry (GC-EI-MS), this approach has been established since the 1970s; the NIST GC-MS library contains reference spectra for nearly 350,000 compounds. Liquid chromatography (LC)-ESI-MS/MS reached widespread use in the 1990s. However, instruments and measurement conditions are much more variable and only in the last decade have comprehensive ESI-MS/MS libraries covering a range of collision energies and fragmentation types become available. NIST MS/MS (https://chemdata.nist.gov/), mzCloud (https://www.mzcloud.org/) and Metlin (https://www.massconsortium.com/) are commercial databases. While valuable and powerful, the cost and/or restrictive usage terms of these resources limit their application and are incompatible with open science and FAIR guidelines, which are gaining importance in the research community. In contrast, multiple community-driven resources have emerged to provide spectral library resources with free usage terms. MassBank (https://massbank.eu) was originally developed in Japan as a federated library for mass spectra, providing both a web platform for spectrum deposition and a specified format and requirements for metadata,[1] and is currently maintained as part of the NFDI4Chem chemical infrastructure (https://www.nfdi4chem.de/). GNPS combines spectral library infrastructure with an ecosystem for online data processing, offering multiple workflows including molecular networking.[2] GNPS data deposition has minimal metadata requirements and a low barrier of entry, bringing together spectral library data from various sources and quality. MassBank of North America (MoNA, https://massbank.us) collects data from user contributions and different sources, including MassBank and GNPS, and serves search functionality with APIs.

One advantage of commercial libraries is that they are extensively curated by hand to ensure high data quality. In contrast, contributors to community-sourced libraries can rarely dedicate a lot of resources to detailed curation. As a result, there is a tradeoff between library curation workload, spectral quality, and library size. With the emergence of non-target screening (NTS) in environmental analytical chemistry, in 2010 Eawag joined a community effort by the NORMAN Network (https://www.norman-network.net/) to populate MassBank with spectra of environmental micropollutants, previously scarcely represented in LC-ESI-MS/MS libraries.[3] Up to 28 spectrum types per compound (including varying collision energy, polarity, resolution) were acquired to cover a range of measurement conditions. To keep curation efforts manageable while ensuring high data quality, we developed the automated curation workflow RMassBank.[4] Briefly, RMassBank extracts ion chromatogram traces and matching MS2 spectra for specified analytes, annotates putative formulas to fragments, and uses this information to build a recalibration curve. After a second more restrictive formula annotation, unmatched peaks as well as spurious fragments are discarded using customizable filter criteria. Furthermore, chemical metadata is collected from online databases and added to the record after manual review (Fig. 2). While initially intended as an in-house tool, it was adopted by MassBank contributors from NORMAN and beyond. This workflow facilitated the establishment of MassBank as the prime open platform for high-quality HRMS reference spectra, now containing >100k spectra for >16k compounds, predominantly from Orbitrap and QToF instruments.

RMassBank is still under continued development. In addition to filtering by formula match, a combined fragment chromatogram correlation and statistical approach was introduced, improving the processed spectrum quality for large (>1,000 Da) molecules and compounds with fluorine atoms, which frequently caused issues with fragment annotation and noise filtering. Further, a viewer was introduced for manual review and spectrum selection without a full MassBank installation. In addition to its main purpose for library autocuration, RMassBank has also been applied as a general purpose MS2 extraction and preprocessing tool in NTS workflows.[5-9]

In the context of automatic curation, ideas from RMassBank are echoed in later approaches, such as WEIZMASS,[10] LibGen[11] and MSMS-Chooser.[12] Features like self-recalibration have been implemented in commercial products as well, *e.g.* the mzVault library generation tool accompanying Compound Discoverer (Thermo Scientific, Bremen). Brungs *et al.* introduced an autocuration approach based on mzmine (https://github.com/mzmine/mzmine), focused on high throughput and reproducibility.[13]

MassBank and GNPS constitute the main source of openly available spectral libraries. Together, the resources contain spectra for nominally over 34,000 individual chemical structures. When considering only high-resolution spectra and conducting further dereplication and quality filtering, the 'open dataset' consists of around 15,000 unique structures with 200,000 spectra. The Brungs autocuration method is now being applied to recently acquired spectra for 20k compounds, generating >1 million library spectra from >16k standards with detailed metadata.[13] Together with MassBank und GNPS spectral libraries, these large-scale efforts are expanding the size and quality of the open spectral library dataset markedly and represent an important step forward in the field.

A comment on the limited meaning of raw numbers in this context: For the usefulness of a library, the content type and diversity are often more relevant than the raw compound count. Library size can efficiently be inflated by including large numbers of di/tripeptides that exhibit very systematic fragmentation patterns and also have little relevance to real-life measurements. As a dataset for machine learning, diversity and chemical coverage is important; *e.g.* a library of 10,000 tripeptides with quite systematic fragmentation provides arguably less information than a library of 2000 diverse small molecules. For direct use as search libraries, scope is relevant: A library with 10,000 natural products is of limited use to an environmental chemist, and *vice versa*, a library with 10,000 pharmaceuticals is of little use to a plant metabolomics practitioner. Finally, even combining open and commercial datasets, spectral libraries still cover a small fraction of

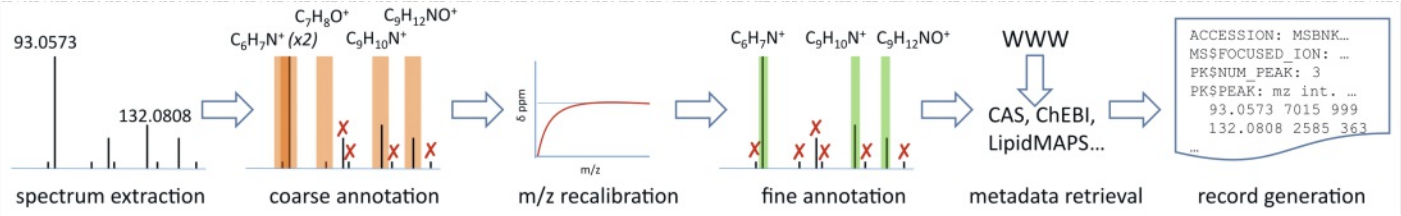

Fig. 2. RMassBank, an automated spectrum library generation and curation workflow. First, raw spectra are extracted for target compounds. Subsequently, peaks are annotated by subformula first coarsely to generate a recalibration curve, and then more narrowly to filter out invalid peaks. Metadata is retrieved from online sources and a MassBank record ready for upload is created.

known molecules and their known and unknown transformation products.

2.2 *Spectral Libraries in Machine Learning*

In addition to facilitating nontarget analysis, spectral databases constitute a central resource for training and evaluation of CompMS algorithms. Legacy algorithms were usually trained on datasets obtained through specific collaborations or potentially through means such as web scraping, making the results difficult to reproduce and compare. MassBank, MoNA, and GNPS make training data freely available with clear terms of usage and make it possible to conduct machine learning research in a reproducible manner. While this is a significant advance, the practical usability of models trained on this data alone can vary. The best performing approaches trusted in practical applications also include data extracted from the NIST MSMS library for training and evaluation, totaling approximately 35,000 unique structures with 1,500,000 spectra. Since this data cannot be freely distributed, it is currently necessary to train and/or evaluate algorithms on open data (GNPS + MassBank/MoNA) and a complete dataset (including also NIST) separately. As noted above, the openly available data pool continues to rapidly expand; with data for >20k new structures now available, significant progress for purely open-data models can be expected in the upcoming years. One continuing challenge is the evaluation of CompMS methods. Partly due to dataset restrictions, there is currently no broadly accepted standard for method evaluation. The Critical Assessment of Small Molecule Identification (CASMI) challenges (http://casmi-contest.org) have periodically evaluated methods on truly novel data with predefined metrics to assess the progress of the field, but the focus and rules have differed over editions. To address this gap, efforts are underway to create a comprehensive dataset and benchmark for the most important machine learning tasks in CompMS. We aim to incorporate both existing and new open datasets and to make this data easily accessible to machine learning practitioners with minimal barriers to entry (Bushuiev *et al.* in preparation).

3. Machine Learning Applications

3.1 *Spectrum Prediction*

An intuitive application of computational methods to mass spectrometry is spectrum prediction. Predicted spectra can augment spectral libraries and broaden the chemical space addressable by spectral library searches from a mere 10,000's of measured reference spectra to any desired number of known structures.

Originally, rule-based approaches such as EPIC,[14] MetFrag,[15] and MAGMA[16,17] were developed, which predicted only fragment masses but no associated intensities ('barcode spectra'). These approaches are unselective and generate a large amount of false positive peaks. Current research concentrates on the prediction of complete spectra, *i.e.* peaks with associated intensities; applying different deep learning methods from simple feed-forward neural networks to transformers and 3D graph convolutions. One approach is prediction which is based on fragmentation mechanisms. This includes the seminal CFM-ID and its later iterations[18-21] as well as the recent ICEBERG model by Goldman *et al.*[22] In contrast, some approaches predict binned spectra, where intensities for every mass unit are predicted without an explicit fragmentation model, usually ignoring high-resolution accurate mass. This includes NEIMS[23] (which was originally trained for low-resolution EI spectra), MassFormer[24] and work by Zhu *et al.*[25] and Hong *et al.*[26] A third approach which leverages high-resolution data without requiring an explicit fragmentation model is the prediction of formula-intensity pairs. This approach is taken by GRAFF-MS,[27] RASSP[28] and SCARF.[29] Some spectral representations take exact mass into account without assigning formulas or structures to fragments;[30] these may serve as a basis for future spectra prediction methods.

Despite the diversity of approaches currently being developed, spectrum prediction remains a challenging task. State-of-the-art methods reach average spectrum similarities of ~0.7, with 1.0 being a perfect match and 0.7–0.8 being typical cutoffs applied in practice. However, this number is prone to artifacts. When looking at structure retrieval (returning the correct candidate in a search), current methods achieve top-1 retrieval in only 10–20% of cases, highlighting that practical application is still limited.

3.2 *Molecular Fingerprint Prediction*

Although spectral prediction is an intuitive approach to compound annotation, other methods exist to match experimental spectra to candidate molecules. In molecular fingerprint prediction, the reverse approach is taken: a model is trained to predict chemical properties and structural features from spectra. The feature vector predicted for experimental spectra is then used for similarity search in a database of compounds (with the same precursor, in the case of ESI-MS). For GC-EI-MS, classifiers for presence/absence of 70 chemical features were developed in the 1990s[31] and later applied in combination with chemical structure generation.[32] However, exhaustive structure generation quickly becomes infeasible even at low masses. For ESI-MS, FingerID[33] established the approach of training a support vector machine (SVM) to predict molecular fingerprints (long bit vectors of presence/absence of chemical structural features) and subsequently matching predicted fingerprints against a chemical structure database. The approach was initially limited by available training data. CSI:FingerID[34] expanded on the method, combining multiple kernels based on mass spectral features and fragmentation trees,[35] and provided predictions for 1415 fingerprint features. CSI:FingerID has been subject to extensive development not only on the algorithm but also with regard to software engineering.[36] It currently predicts 5,000 fingerprint bits for both positive-mode and negative-mode spectra.

Alternative approaches for fingerprint prediction have recently been developed. MetFID[37] and IDSL_MINT[38] apply neural networks instead of SVMs to predict molecular fingerprints. MIST[39] uses a transformer architecture using chemical formulas as input, fine-tuned with contrastive learning to improve annotation accuracy. Dührkop *et al.* recently combined kernel methods with neural networks to improve fingerprint prediction in

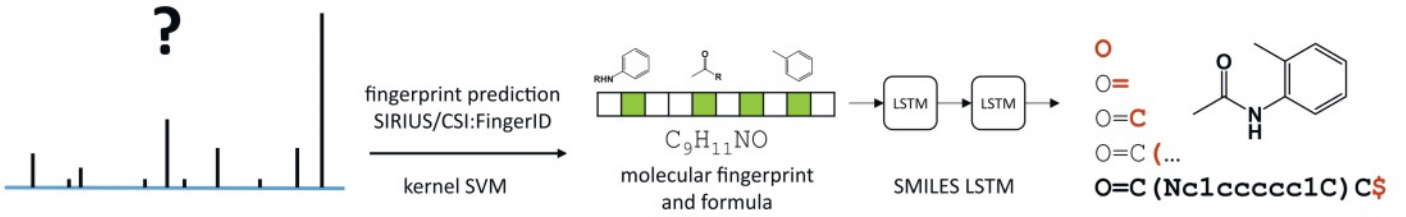

Fig. 3. High-level overview of MSNovelist. In the first step, SIRIUS / CSI:FingerID predict the molecular fingerprint and formula from an unknown spectrum. In the second step, a SMILES code encoding the molecular structure is generated by repeated (autoregressive) application of a long-short term memory (LSTM) recurrent neural network.

CSI:FingerID.[40] While these methods appear to slightly improve fingerprint prediction accuracy, top-1 retrieval among the different approaches remains roughly comparable. On a 3,868-compound reference dataset from GNPS, CSI:FingerID reaches top-1 retrieval in 38% of cases. It must be taken into account that this is a search against a database with tens of millions of compounds, designed to be challenging for the purpose of characterizing method performance.[41] In practice, the search space can be restricted to plausible candidates, such as molecules from biological databases (approximately 1 million structures[42]) or with some exposome-related metadata (approximately 400,000 structures[43]), substantially increasing the success of these methods in practice.

Interestingly, predicting molecular fingerprints from mass spectra opens the door for further applications. Predicted fingerprints can themselves be used as input for machine learning methods to predict other chemical properties, allowing for a prediction of spectrum-to-fingerprint-to-X (for example, toxicity, ionization, chemical classes) without determining the specific chemical structure. The downstream model can be trained with (theoretically) any available dataset for which the fingerprints can be computed. This method allows for concentration estimation[44] or toxicity prediction[45,46] for LC-HRMS features without identification. CANOPUS[42] was trained to predict chemical classes from MS2 spectra, allowing a broader chemical picture of a sample to be obtained without identification. Importantly, it highlighted the fact that large fingerprint-to-X training sets can be computed from chemical databases for deterministic properties.

3.3 *MSNovelist: The Journey Towards De Novo Structure Prediction*

In silico database search methods make it possible to search for structures with unknown spectra. However, they fundamentally cannot search for novel, truly unknown structures. This limits the applicability in various applications, such as identifying environmental transformation products, drug metabolites or novel natural products. To overcome this limitation, a method would be required to generate completely new structures rather than looking up structures in a database. Given the very large amount of possible chemical structures and the limited information available from MS2 spectra, this is a very challenging task. First attempts were made with combinatorial approaches for GC-MS[32] and LC-MS,[47] but struggled with combinatorial explosion due to the vast chemical space accessible by such methods. However, a new avenue towards *de novo* structure prediction emerged from recent molecule generation methods based on deep learning. These approaches conditionally (rather than randomly or exhaustively) generate molecules and therefore directly access molecules with desired properties rather than generating extremely large sets of molecules and filtering post-hoc. The approaches range from generation of molecules as SMILES strings (a line notation of chemical structure), generation of graphs or simplified graph representations, to diffusion methods.

The core idea of MSNovelist is the combination of advances in molecular fingerprint prediction with emerging molecule generation methods. It was postulated that structural information from predicted molecular fingerprints could be used to condition molecule generation. Crucially, we built on the insight from Dührkop *et al.*[42] that large training sets for deterministic data can be generated from chemical databases. While there are only 10,000s of training examples for spectra-to-fingerprint prediction, millions can be generated for fingerprint-to-structure prediction. The MSNovelist model consists of a fingerprint-to-SMILES decoder implemented as a long-short term memory (LSTM) neural network,[48] where a SMILES string is generated token by token by recursively applying the decoder model (Fig. 3). The SMILES LSTM is similar to the decoder of autoencoder models;[49] appropriate feature engineering guides the generation of molecules of a specified formula. In contrast to stochastic sampling strategies used in models for drug design, which results in a diverse set of generated molecules, MSNovelist applies beam search[50] to approximate the highest-probability SMILES sequences overall.

The model was evaluated on the GNPS dataset also used for CSI:FingerID and reached a 26% top-1 recovery (compared to 38% for CSI:FingerID). For spectra correctly annotated by CSI:FingerID a top-1 recovery of 64% was achieved. Although the annotation rates were low, this model demonstrated for the first time that *de novo* structure prediction is in fact feasible. Evaluation of *de novo* models is intrinsically difficult. While in database search evaluation the search space is clearly delineated, it is unclear what a desirable chemical space for *de novo* generation should be. A model performing highly in evaluation, reproducing known chemistry very well, might not be able to find truly new molecules outside of the known chemical space. In the context of MSNovelist, we used fingerprint similarities as surrogates to show that the model systematically improved over the training set; however, this evaluation is specific to MSNovelist's use of a fingerprint intermediate and is not applicable as an evaluation method for every potential *de novo* approach.

Multiple efforts to apply deep learning-driven generative chemistry to *de novo* structure prediction were undertaken in parallel to or following MSNovelist. MassGenie[51] is a transformer-based *de novo* model pre-trained on *in silico* barcode spectra and refined on GNPS data. Interestingly, the model works purely on m/z values and ignores intensities except as a cutoff criterion. Spec2Mol[52] is a *de novo* method combining a SMILES autoencoder and a convolutional neural network to predict the autoencoder embedding. Mass2SMILES[53] combines a transformer encoder with a temporal convolutional network to generate SMILES from spectra input. Kutuzova *et al.* developed a bimodal spectrum-structure embedding, enabling prediction of both spectra from structure and structure from spectra.[54] MS2Mol is a transformer model predicting both molecular formula and structure *de novo* from MS2 spectra.[55] This surge of publications after a decade of inactivity shows increased interest in the field. However, the models are hard to compare because of the diverse evaluation methods used. In summary, despite progress, the complexity of small molecule chemistry continues to make *de novo* structure prediction challenging.

4. Perspectives and Outlook

Herein, I briefly reviewed important developments and challenges in CompMS and embedded our work in this context. It should be noted that I only covered work directly related to classical compound identification; further, I considered molecular formula determination a sufficiently solved problem,[56,57] which may not be the case in practice. An additional prominent research topic in the recent literature is the characterization of structural similarity (rather than compound identification) based on MS2 spectra. This topic includes molecular networking, *i.e.* building networks of related features based on spectral similarity or related ions,[2,58] but also spectrum embeddings reflecting structural similarity[59,60] and the use of partial spectral library matches to propagate annotations.[61,62] Over the last decade, CompMS has matured and expanded significantly. Academic proof-of-concept work has developed into mainstream software, and more research groups have started working on CompMS topics. Fingerprint prediction is now an established method and has already undergone significant development, such that the improvement by new approaches is of limited magnitude. In contrast, the topic of *de novo* elucidation and a variety of deep learning applications are showing new avenues. CompMS remains a medium-data regime, where sufficient but not abundant labeled data is available for training, so that mass spectrometry is still very incompletely described by available data. Semi-supervised methods attempt to incorporate the larger amount of unlabeled data available.[63] In cheminformatics in general, recent chemical foundation models incorporate chemical information across multiple modalities,[64] *e.g.* for chemical property prediction. Future work may consider CompMS tasks not in isolation, but as one of many outputs for multimodal and/or foundation models. By combining chemical information and diverse knowledge sources, it may be possible to capture chemical relations and mechanisms not represented in pure CompMS training data, enabling a further leap in model quality.

Acknowledgements

I would like to thank Dr. Jennifer Schollée, Prof. Dr. Juliane Hollender and Prof. Dr. Nicola Zamboni for their helpful comments on the manuscript. I thank the attendees of the Dagstuhl Seminar Series on Computational Metabolomics for scientific exchange and inspiration. I acknowledge Eawag and ETH Zürich for funding. The wording of a few individual sentences was revised with use of artificial intelligence.

Author Contributions

M.S researched literature and wrote and edited the manuscript.

Received: May 30, 2024

[1] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, *J. Mass Spectrom.* **2010**, *45*, 703, https://doi.org/10.1002/jms.1777.

[2] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, N. Bandeira, *Nat. Biotechnol.* **2016**, *34*, 828, https://doi.org/10.1038/nbt.3597.

[3] T. Schulze, E. Schymanski, M. Stravs, S. Neumann, M. Krauss, H. Singer, C. Hug, C. Gallampois, J. Hollender, J. Slobodnik, W. Brack, *NORMAN Netw. Bull.* **2012**, *3*, 9.

[4] M. A. Stravs, E. L. Schymanski, H. P. Singer, J. Hollender, *J. Mass Spectrom.* **2013**, *48*, 89, https://doi.org/10.1002/jms.3131.

[5] M. A. Stravs, F. Pomati, J. Hollender, *Environ. Sci. Process. Impacts* **2017**, *19*, 822, https://doi.org/10.1039/C7EM00100B.

[6] K. Kiefer, L. Du, H. Singer, J. Hollender, *Water Res.* **2021**, *196*, 116994, https://doi.org/10.1016/j.watres.2021.116994.

[7] V. Albergamo, J. E. Schollée, E. L. Schymanski, R. Helmus, H. Timmer, J. Hollender, P. de Voogt, *Environ. Sci. Technol.* **2019**, *53*, 7584, https://doi.org/10.1021/acs.est.9b01750.

[8] S. L. Rich, D. E. Helbling, *Environ. Sci. Technol.* **2023**, *57*, 10404, https://doi.org/10.1021/acs.est.3c02408.

[9] J. E. Schollée, E. L. Schymanski, S. E. Avak, M. Loos, J. Hollender, *Anal. Chem.* **2015**, *87*, 12121, https://doi.org/10.1021/acs.analchem.5b02905.

[10] N. Shahaf, I. Rogachev, U. Heinig, S. Meir, S. Malitsky, M. Battat, H. Wyner, S. Zheng, R. Wehrens, A. Aharoni, *Nat. Commun.* **2016**, *7*, 12423, https://doi.org/10.1038/ncomms12423.

[11] F. Kong, U. Keshet, T. Shen, E. Rodriguez, O. Fiehn, *Anal. Chem.* **2023**, *95*, 16810, https://doi.org/10.1021/acs.analchem.3c02263.

[12] F. Vargas, K. C. Weldon, N. Sikora, M. Wang, Z. Zhang, E. C. Gentry, M. W. Panitchpakdi, A. M. Caraballo-Rodríguez, P. C. Dorrestein, A. K. Jarmusch, *Rapid Commun. Mass Spectrom.* **2020**, *34*, e8725, https://doi.org/10.1002/rcm.8725.

[13] Efficient Generation of Open Multi-Stage Fragmentation Mass Spectral Libraries, C. Brungs, R. Schmid, S. Heuckeroth, A. Mazumdar, M. Drexler, P. Šácha, P. C. Dorrestein, D. Petras, L.-F. Nothias, R. Nencka, Z. Kameník, T. Pluskal, *ChemRxiv*, **2024**, https://doi.org/10.26434/chemrxiv-2024-l1tqh.

[14] A. W. Hill, R. J. Mortishire-Smith, *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111, https://doi.org/10.1002/rcm.2177.

[15] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, *BMC Bioinformatics* **2010**, *11*, 148, https://doi.org/10.1186/1471-2105-11-148.

[16] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, *Mass Spectrom.* **2014**, *3*, S0033, https://doi.org/10.5702/massspectrometry.S0033.

[17] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, J. Vervoort, *Rapid Commun. Mass Spectrom.* **2012**, *26*, 2461, https://doi.org/10.1002/rcm.6364.

[18] F. Allen, R. Greiner, D. Wishart, *Metabolomics* **2015**, *11*, 98, https://doi.org/10.1007/s11306-014-0676-4.

[19] F. Allen, A. Pon, M. Wilson, R. Greiner, D. Wishart, *Nucleic Acids Res.* **2014**, *42*, 94, https://doi.org/10.1093/nar/gku436.

[20] Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, D. S. Wishart, *Metabolites* **2019**, *9*, 72, https://doi.org/10.3390/metabo9040072.

[21] F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, D. S. Wishart, *Anal. Chem.* **2021**, *93*, 11692, https://doi.org/10.1021/acs.analchem.1c01465.

[22] S. Goldman, J. Li, C. W. Coley, *Anal. Chem.* **2024**, *96*, 3419, https://doi.org/10.1021/acs.analchem.3c04654.

[23] J. N. Wei, D. Belanger, R. P. Adams, D. Sculley, *ACS Cent. Sci.* **2019**, *5*, 700, https://doi.org/10.1021/acscentsci.9b00085.

[24] A. Young, H. Röst, B. Wang, *Nat. Mach. Intell.* **2024**, *6*, 404, https://doi.org/10.1038/s42256-024-00816-8.

[25] H. Zhu, L. Liu, S. Hassoun, *ArXiv* 201004661 Cs **2020**.

[26] Y. Hong, S. Li, C. J. Welch, S. Tichy, Y. Ye, H. Tang, *Bioinformatics* **2023**, *39*, btad354, https://doi.org/10.1093/bioinformatics/btad354.

[27] M. Murphy, S. Jegelka, E. Fraenkel, T. Kind, D. Healey, T. Butler, 'Efficiently predicting high resolution mass spectra with graph neural networks', https://arxiv.org/abs/2301.11419v1, accessed May 9, 2023.

[28] R. L. Zhu, E. Jonas, *Anal. Chem.* **2023**, *95*, 2653, https://doi.org/10.1021/acs.analchem.2c02093.

[29] S. Goldman, J. Bradshaw, J. Xin, C. Coley, *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 48548.

[30] T. Altenburg, S. Wang, T. Muth, B. Y. Renard, *bioRxiv* **2020**, 2020.05.19.101345, https://doi.org/10.1101/2020.05.19.101345.

[31] K. Varmuza, W. Werther, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323, https://doi.org/10.1021/ci9501406.

[32] E. L. Schymanski, C. Meinert, M. Meringer, W. Brack, *Anal. Chim. Acta* **2008**, *615*, 136, https://doi.org/10.1016/j.aca.2008.03.060.

[33] H. Shen, N. Zamboni, M. Heinonen, J. Rousu, *Metabolites* **2013**, *3*, 484, https://doi.org/10.3390/metabo3020484.

[34] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. (USA)* **2015**, *112*, 12580, https://doi.org/10.1073/pnas.1509788112.

[35] F. Rasche, A. Svatos, R. K. Maddula, C. Böttcher, S. Böcker, *Anal. Chem.* **2011**, *83*, 1243, https://doi.org/10.1021/ac101825k.
[36] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, S. Böcker, *Nat. Methods* **2019**, *16*, 299, https://doi.org/10.1038/s41592-019-0344-8.
[37] Z. Fan, A. Alley, K. Ghaffari, H. W. Ressom, *Metabolomics* **2020**, *16*, 104, https://doi.org/10.1007/s11306-020-01726-7.
[38] S. F. Baygi, D. K. Barupal, *J. Cheminformatics* **2024**, *16*, 8, https://doi.org/10.1186/s13321-024-00804-5.
[39] S. Goldman, J. Wohlwend, M. Stražar, G. Haroush, R. J. Xavier, C. W. Coley, *Nat. Mach. Intell.* **2023**, *5*, 965, https://doi.org/10.1038/s42256-023-00708-3.
[40] K. Dührkop, *Bioinformatics* **2022**, *38*, i342, https://doi.org/10.1093/bioinformatics/btac260.
[41] S. Böcker, *Curr. Opin. Chem. Biol.* **2017**, *36*, 1, https://doi.org/10.1016/j.cbpa.2016.12.010.
[42] K. Dührkop, L.-F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein, S. Böcker, *Nat. Biotechnol.* **2020**, *1*, https://doi.org/10.1038/s41587-020-0740-8.
[43] E. L. Schymanski, T. Kondić, S. Neumann, P. A. Thiessen, J. Zhang, E. E. Bolton, *J. Cheminformatics* **2021**, *13*, 19, https://doi.org/10.1186/s13321-021-00489-0.
[44] H. Sepman, L. Malm, P. Peets, M. MacLeod, J. Martin, M. Breitholtz, A. Kruve, *Anal. Chem.* **2023**, https://doi.org/10.1021/acs.analchem.3c01744.
[45] I. Rahu, M. Kull, A. Kruve, *J. Chem. Inf. Model.* **2024**, *64*, 3093, https://doi.org/10.1021/acs.jcim.3c02050.
[46] K. Arturi, J. Hollender, *Environ. Sci. Technol.* **2023**, *57*, 18067, https://doi.org/10.1021/acs.est.3c00304.
[47] J. E. Peironcely, M. Rojas-Chertó, A. Tas, R. Vreeken, T. Reijmers, L. Coulier, T. Hankemeier, *Anal. Chem.* **2013**, *85*, 3576, https://doi.org/10.1021/ac303218u.
[48] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735, https://doi.org/10.1162/neco.1997.9.8.1735.
[49] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268, https://doi.org/10.1021/acscentsci.7b00572.
[50] A. Graves, 'Sequence Transduction with Recurrent Neural Networks' *arXiv*:1211.3711, *arXiv*, **2012**, https://doi.org/10.48550/arXiv.1211.3711.
[51] A. D. Shrivastava, N. Swainston, S. Samanta, I. Roberts, M. W. Muelas, D. B. Kell, *bioRxiv* **2021**, 2021.06.25.449969, https://doi.org/10.1101/2021.06.25.449969.
[52] E. E. Litsa, V. Chenthamarakshan, P. Das, L. E. Kavraki, *Commun. Chem.* **2023**, *6*, 1, https://doi.org/10.1038/s42004-023-00932-3.
[53] D. Elser, F. Huber, E. Gaquerel, *bioRxiv* **2023**, 2023.07.06.547963, https://doi.org/10.1101/2023.07.06.547963.
[54] S. Kutuzova, C. Igel, M. Nielsen, D. McCloskey, 'Bi-Modal Variational Autoencoders for Metabolite Identification Using Tandem Mass Spectrometry', *bioRxiv*, **2021**, https://doi.org/10.1101/2021.08.03.454944.
[55] T. Butler, A. Frandsen, R. Lightheart, B. Bargh, T. Kerby, K. West, J. Davison, J. Taylor, C. Krettler, T. Bollerman, G. Voronov, K. Moon, T. Kind, P. Dorrestein, A. Allen, V. Colluru, D. Healey, 'MS2Mol: A transformer model for illuminating dark chemical space from mass spectra', *ChemRxiv*, **2023**, https://doi.org/10.26434/chemrxiv-2023-vsmpx-v4.
[56] K. Dührkop, K. Scheubert, S. Böcker, *Metabolites* **2013**, *3*, 506, https://doi.org/10.3390/metabo3020506.
[57] S. Goldman, J. Xin, J. Provenzano, C. W. Coley, *J. Chem. Inf. Model.* **2023**, https://doi.org/10.1021/acs.jcim.3c01082.
[58] R. Schmid, D. Petras, L.-F. Nothias, M. Wang, A. T. Aron, A. Jagels, H. Tsugawa, J. Rainer, M. Garcia-Aloy, K. Dührkop, A. Korf, T. Pluskal, Z. Kameník, A. K. Jarmusch, A. M. Caraballo-Rodríguez, K. C. Weldon, M. Nothias-Esposito, A. A. Aksenov, A. Bauermeister, A. Albarracin Orio, C. O. Grundmann, F. Vargas, I. Koester, J. M. Gauglitz, E. C. Gentry, Y. Hövelmann, S. A. Kalinina, M. A. Pendergraft, M. Panitchpakdi, R. Tehan, A. Le Gouellec, G. Aleti, H. Mannochio Russo, B. Arndt, F. Hübner, H. Hayen, H. Zhi, M. Raffatellu, K. A. Prather, L. I. Aluwihare, S. Böcker, K. L. McPhail, H.-U. Humpf, U. Karst, P. C. Dorrestein, *Nat. Commun.* **2021**, *12*, 3832, https://doi.org/10.1038/s41467-021-23953-9.
[59] F. Huber, S. van der Burg, J. J. J. van der Hooft, L. Ridder, *J. Cheminformatics* **2021**, *13*, 84, https://doi.org/10.1186/s13321-021-00558-4.
[60] F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers, J. J. J. van der Hooft, *PLOS Comput. Biol.* **2021**, *17*, e1008724, https://doi.org/10.1371/journal.pcbi.1008724.
[61] R. R. da Silva, M. Wang, L.-F. Nothias, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes, P. C. Dorrestein, *PLOS Comput. Biol.* **2018**, *14*, e1006089, https://doi.org/10.1371/journal.pcbi.1006089.
[62] W. Bittremieux, N. E. Avalon, S. P. Thomas, S. A. Kakhkhorov, A. A. Aksenov, P. W. P. Gomes, C. M. Aceves, A. M. Caraballo-Rodríguez, J. M. Gauglitz, W. H. Gerwick, T. Huan, A. K. Jarmusch, R. F. Kaddurah-Daouk, K. B. Kang, H. W. Kim, T. Kondić, H. Mannochio-Russo, M. J. Meehan, A. V. Melnik, L.-F. Nothias, C. O'Donovan, M. Panitchpakdi, D. Petras, R. Schmid, E. L. Schymanski, J. J. J. van der Hooft, K. C. Weldon, H. Yang, S. Xing, J. Zemlin, M. Wang, P. C. Dorrestein, *Nat. Commun.* **2023**, *14*, 8488, https://doi.org/10.1038/s41467-023-44035-y.
[63] R. Bushuiev, A. Bushuiev, R. Samusevich, J. Šivic, T. Pluskal, *ChemRxiv* **2023**, https://doi.org/10.26434/chemrxiv-2023-kss3r.
[64] J. Chang, J. C. Ye, *Nat. Commun.* **2024**, *15*, 2323, https://doi.org/10.1038/s41467-024-46440-3.