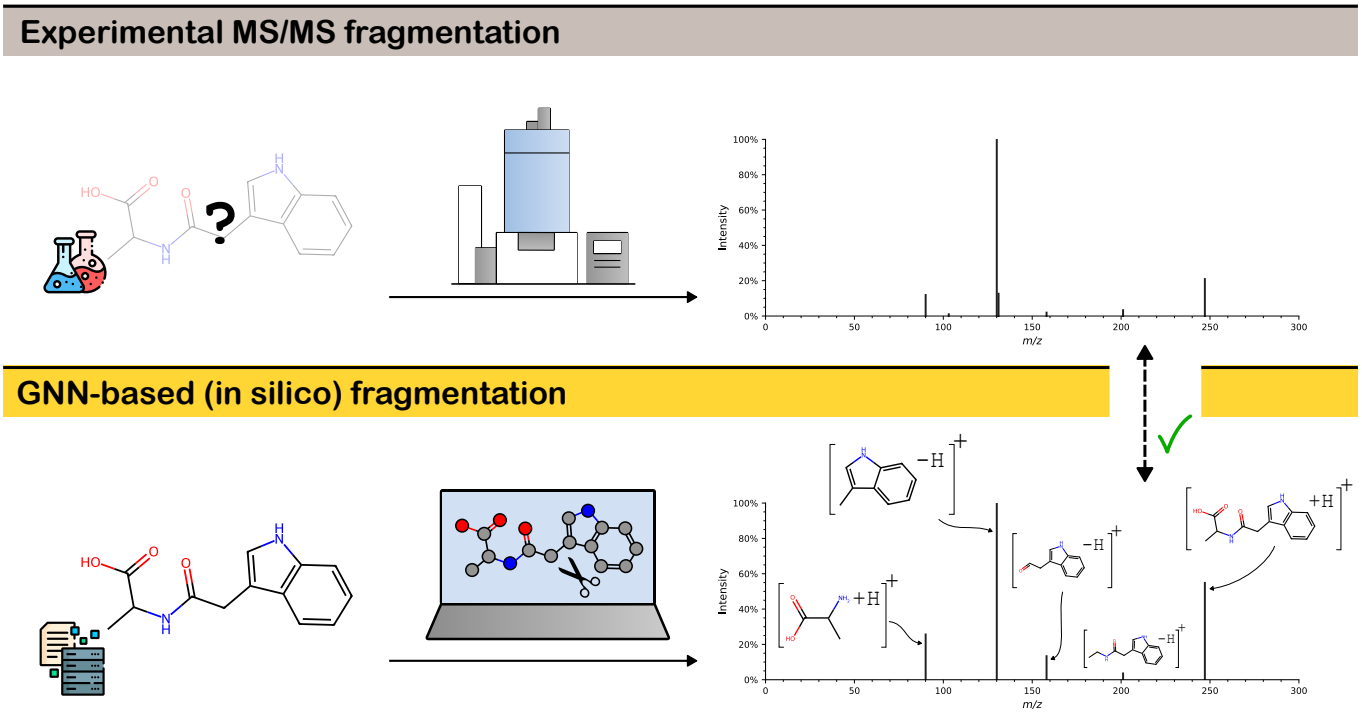

FIORA: Local neighborhood-based prediction of compound mass spectra from single fragmentation events

Yannek Nowatzky [1], Francesco Russo,[1] Jan Lisec [1], Alexander Kister,[1] Knut Reinert [2], Thilo Muth [2, 3] and Philipp Benner [1,*]

[1] Federal Institute for Materials Research and Testing (BAM), Berlin, Germany
[2] Freie Universität Berlin, Berlin, Germany
[3] Robert Koch Institute, Berlin, Germany
[*] Lead Contact and correspondence: philipp.benner@bam.de

ABSTRACT

Non-targeted metabolomics holds great promise for advancing precision medicine and facilitating the discovery of novel biomarkers. However, the identification of compounds from tandem mass spectra remains a non-trivial task due to the incomplete nature of spectral reference libraries. Augmenting these libraries with simulated mass spectra can provide the necessary reference to resolve unmatched mass spectra, but remains a difficult undertaking to this day. In this study, we introduce FIORA, an innovative open-source algorithm using graph neural networks to simulate tandem mass spectra *in silico*. Our objective is to improve fragment intensity prediction with an intricate graph model architecture that facilitates edge prediction, thereby modeling fragment ions as the result of singular bond breaks and their local molecular neighborhood. We evaluate the performance on test data from NIST (2017) and the curated MS-Dial spectral library, as well as compounds from the 2016 and 2022 CASMI challenges. FIORA not only surpasses state-of-the-art fragmentation algorithms, ICEBERG and CFM-ID, in terms of prediction quality, but also predicts additional features, such as retention time and collision cross section. In addition, FIORA demonstrates significant speed improvements through the use of GPUs. This enables rapid (re)scoring of putative compound identifications in non-targeted experiments and facilitates large-scale expansion of spectral reference libraries with accurate spectral predictions.

MAIN

Introduction

Progress in non-targeted metabolomics is limited by the scarcity of high-quality reference spectra. This discipline promotes an unbiased exploration of metabolites within biological systems and is facilitated by liquid chromatography-mass spectrometry (LC-MS) [1]. In high-throughput settings, compounds are ionized and isolated based on their biophysical properties and ion mass, and then fragmented into product ions [2, 3]. The product ions are recorded as peaks in a tandem mass spectrum (MS/MS) and act as a signature or fingerprint for a given molecule. However, in 2015, da Silva *et al.* showed that only a small fraction of MS/MS spectra from non-targeted experiments can be annotated by searching spectral libraries of reference standards due to their incomplete nature [4]. They coined the term "dark matter" to describe the overwhelming number of unidentified signals and chemical species that remain unknown.

In the past decade, this situation has led to the development of various algorithms that attempt to infer compound identity directly from mass spectra, so-called *in silico* methods. These include, but are not limited to, CSI:FingerID as part of the SIRIUS suite [5, 6], MS-FINDER [7, 8], and MS2LDA [9]. Despite these advancements, the identification rates of "unknown" compounds remain low. This was demonstrated in the 2016 CASMI challenge, where *in silico* methods achieved a recall rate of only up to 34%, when annotating spectra of previously unknown compounds [10]. Meanwhile, in the most recent CASMI challenge in 2022, identification rates were below 30% [11].

To improve compound identification by amending existing libraries, many research groups attempt to build theoretical product ion spectra from molecular structures [12]. These spectra serve as reference and allow the expansion of public spectral libraries when experimental metabolite spectra are unavailable. In particular, *in silico* fragmentation algorithms simulate the MS fragmentation process, and exploit the chemical and structural properties of the molecules to predict fragment ions or infer their identity. This process is facilitated by large knowledge bases that provide known chemical structures and properties, such as PubChem [13] and HMDB [14]. These are orders of magnitudes larger than spectral databases [15], such as GNPS [16] and METLIN [17]. Nonetheless, the accurate prediction of MS/MS spectra remains a significant challenge due to the scarcity of high-quality training data and algorithms must be thoroughly evaluated to determine their effectiveness for previously unreferenced or unseen metabolites.

At the same time, accurate annotation of compound spectra is paramount to metabolomics. Non-targeted screening approaches

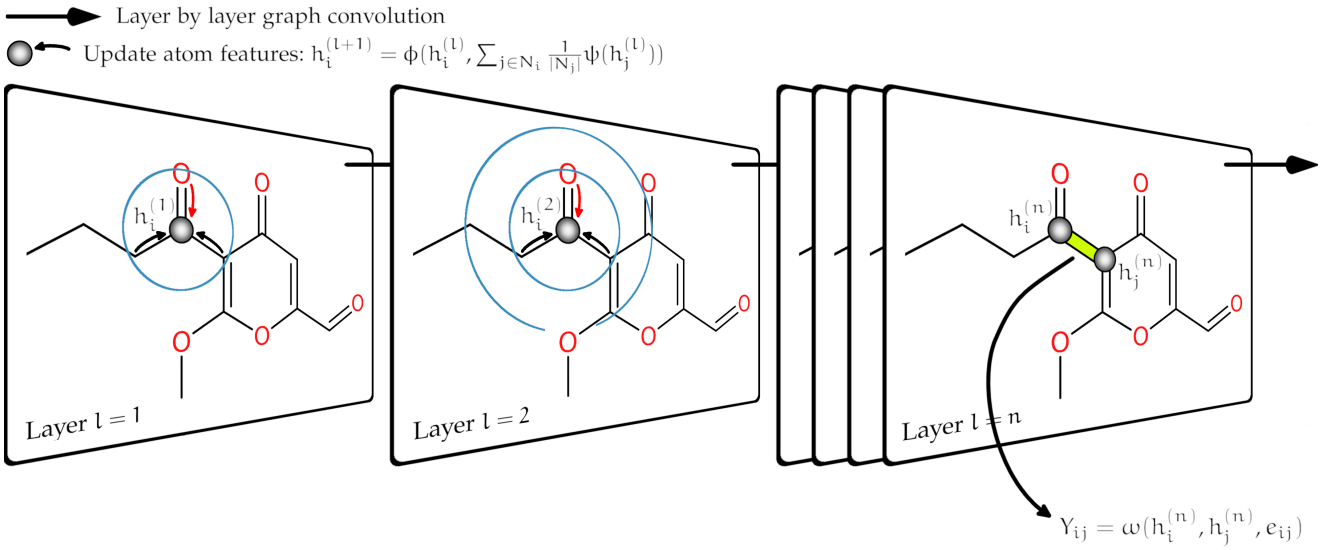

Figure 1: Illustration of how a graph network translates local structure information into molecular property prediction. The network performs multiple graph convolutions on the molecular structure graph, thereby aggregating the local neighborhood, i.e., the surrounding substructure, into hidden representations of the atoms. Subsequently, edge prediction is used to estimate bond properties, such as stability.

have become increasingly popular in clinical diagnostics, drug response monitoring, and the characterization of intracellular molecular mechanisms [18, 19]. The link between the metabolome, which describes the biochemical phenotype, and the genotype, microbiome, and environmental exposures in human health and disease makes metabolomics an invaluable tool for biomarker discovery and hypothesis generation [20]. This places particular emphasis on the development of compound identification methods that must be highly efficient and accurate. Ongoing advances in *in silico* fragmentation methods can deepen our understanding of MS-based compound fragmentation, expand the search space for spectral libraries, and offer additional levels of confidence to other identification methods.

Related work

Bond dissociation is a key concept behind compound fragmentation, as covalent bonds are cleaved during MS/MS, producing fragment ions that appear in the mass spectrum [12, 21]. Typically, one fragment is lost, referred to as neutral loss, while the fragment on the other side of the fragmentation site retains the charge and is observed as a peak in the m/z dimension. Multiple bond cleavages and hydrogen rearrangements may occur. The abundance of the fragment ions and therefore the probabilities of the corresponding bond breaks are directly tied to the peak intensity. *In silico* fragmentation algorithms attempt to identify breakpoints in the molecular structures and use these to impute ion probabilities and peak intensities. The output is a simulated mass spectrum. In addition, structural fragment annotations and fragmentation pathways can be retained.

CFM-ID is an advanced machine learning (ML) algorithm that predicts transition probabilities between fragments. Since its introduction by Allen *et al.* in 2015 [22] it has undergone many improvements, with the latest version 4.0 being published in 2021 [23]. CFM-ID is widely regarded as a pioneer in ML-driven *in silico* fragmentation of molecular structures. The method models the fragmentation process as a stochastic, homogeneous Markov process and learns model parameters using an expectation-maximization algorithm [23]. However, CFM-ID suffers from slow training and prediction performance, often rendering it insufficient for predicting a large candidate space of possible structures or rescoring many tentative identifications. Despite its complexity, CFM-ID does not rely on modern deep learning structures like graph neural networks.

Graph neural networks (GNNs) have become popular over the past decades, in part due to applications in cheminformatics and drug discovery [24]. Their ability to learn and characterize molecular structure graphs has proven to be essential for predicting molecular properties, such as solvation free energy or metabolic stability. While GNNs have recently gained attention in metabolomics applications, their full potential still remains untapped.

One popular approach involves using GNNs to embed the molecular structure and then predict vectors of fixed length representing binned MS/MS spectra. Zhu *et al.* (2020) [25] attempted this using graph convolutional neural networks (GCNs) and graph attention networks (GATs), while Young *et al.* (2021) [26] used a graph transformers architecture [27] for spectra binned at 1 Da resolution. In 2023, Park *et al.* [28] introduced a GNN combining the molecular structure graph with a heterogeneous motif graph, and the QC-GN^{2}oMS^2 model by Overstreet *et al.* (2023) [29] adds quantum chemical bond features to improve spectrum prediction at high resolution. While using modern deep learning architectures, none of these methods leverage the molecular graph structure to their full potential. All the approaches compute a single embedding from the input graph (pooled together from the node features), which is then used to predict a fixed-length vector representing the mass spectrum. This makes the assumption that the models are able to learn all fragment ions from a singular graph representation and directly associate them with the correct m/z bin. In doing so, crucial information about learned subgraph structures around the breaking points becomes entangled when node features are pooled together, making it much harder to learn local properties. This information is accessible at the bond cleavage sites, but requires more complex and elaborate model structures than binned peak prediction. In addition, the fixed output format limits the models to a specific mass resolution, required for binning the spectra. High-resolution predictions get increasingly harder and more training data is required to increase mass accuracy as technology advances. In contrast, fragmentation algorithms that iterate and break chemical bonds allow the direct prediction of fragment ions and the calculation of the exact peak position, so that mass resolution is infinite. Such fragmentation methods are therefore timeless and remain unaffected by the next technological leap in mass specificity.

In 2023, Murphy *et al.* introduced the graph network GRAFF-MS that predicts molecular formulas of fragment ions and neutral

losses from a fixed vocabulary, bypassing the mass resolution problem altogether [30]. They make a case against bond breaking, as most peak signals can be explained by a fixed vocabulary of sufficient length. However, there are several advantages to using bond dissociation for fragment inference. For one, explicitly modeling bond breaks retains a higher level of explainability and allows the annotation of fragments (and fragmentation pathways), which is essential to understanding and validating MS/MS. As all theoretical breaking points are covered, this approach is able to find truly novel fragments in unknown chemical species that may be missing from a fixed vocabulary. Lastly, GRAFF-MS relies on full graph embeddings for molecular formula prediction and, similar to the binned peak prediction models, does not fully exploit the graph substructures around breaking points.

ICEBERG [31] and SCARF [32] are two novel spectral prediction methods developed by Goldman *et al.* (2023) that strike a balance between physically-grounded fragmentation algorithms and the advantages of "black box" peak intensity predictors using deep neural networks. Both models have two separate modules that work in conjunction. The first module generates potential fragments (or molecular formulae in the case of SCARF) and the second module predicts intensities for the set of fragments using Set Transformers [33]. Although these models provide some explanation of possible fragmentation events and predict peaks with high accuracy, they do not take into account the features of the broken bonds or a local representation of their surrounding molecular neighborhood when predicting fragment intensities. We argue that these factors are the most important criteria in determining break probabilities and hydrogen rearrangements. Interestingly, ICEBERG uses GNNs to embed the molecular structure and goes as far as modeling fragmentation events through the stepwise removal of atoms. However, the final intensity prediction trivializes the fragmentation events (and adherent substructures) to the fragment and parent molecule embeddings and the number of bonds removed. Both, ICEBERG and SCARF do not consider covariates, such as collision energy, which is a major factor influencing fragmentation events and, consequently, the resulting fragment ion abundances. Both models operate only in positive ion mode.

Our contribution

We present FIORA, a novel modular network structure that stands for **F**ragment **Io**n **R**econstruction **A**lgorithm. FIORA is a multi-purpose framework designed to predict various spectral features. What sets FIORA apart is the commitment to expressing each bond cleavage with its local molecular neighborhood. This marks a departure from the typical approach of predicting MS/MS spectra or complete sets of fragments based on a summarized representation (embedding) of the molecule as seen in many recent algorithms. Instead, FIORA evaluates bond dissociation events independently, on the basis of their surrounding molecular structure, thereby simulating the physical fragmentation process of MS more directly. FIORA uses state-of-the-art GNN architectures and formalizes fragment ion prediction as an edge-level prediction task within the molecular structure graph. In doing so, FIORA makes great use of high-performance GPUs and has a strong emphasis on explainability in its decision-making process, but is so far limited to single-step fragmentation.

FIORA reconstructs complete MS/MS spectra for both positive and negative ionization modes ([M+H]+ and [M-H]- precursors). In addition, FIORA estimates retention times (RT) and collision cross sections (CCS), which add further dimensions for MS-based compound identification and is a truly novel addition to spectral prediction software. We benchmark the performance against the top-performing methods, CFM-ID and ICEBERG. Our results demonstrate that FIORA learns fragmentation patterns relatively independent of the structural similarity between the training set and unknown compounds. This ensures a high degree of generalizability for modeling truly unknown structures and sets a new state of the art for spectral feature prediction. FIORA is open source (MIT license) and freely available on GitHub at `https://github.com/BAMeScience/fiora`.

RESULTS

Overview of the fragmentation method

The core idea behind FIORA is to predict mass spectra indirectly by anticipating molecular bond breaks that occur during the fragmentation process of tandem MS. To this end, we employ a GNN to learn hidden representations of the molecules and formulate bond breaks as an edge property prediction task, as illustrated in Figure 1. Fragment ions (and neutral losses) are modeled as a direct consequence of edge removal from the molecular graph. Our model takes into account the local neighborhood of each bond, thus exploiting a close-to-complete chemical representation relevant for deciphering fragmentation events and ion rearrangements.

Subsequently, FIORA models MS/MS signals as a probability distribution over the predicted fragment products following a single bond dissociation. It builds upon the statistics of independent break tendency values introduced by Allen *et al.* (2015) [22]. We extend this concept to directly estimate fragment ion probabilities featuring multiple hydrogen rearrangements, as well as estimate the precursor probability. Further details on the model and the spectral reconstruction algorithm can be found in the Methods section.

The graph network module also allows traditional molecular property prediction, as seen in other fields such as drug property prediction [24]. We utilize this for learning RT and CCS values with neural network submodules using the molecular graph embeddings that are a result of the fragmentation process. In this way, FIORA provides multiple MS/MS feature dimensions to match experimental data, which can be used to improve compound identification. To the best of our knowledge, FIORA is the only model that simulates complete MS/MS compound spectra, including fragment annotation, RT, and CCS values. Furthermore, FIORA is designed to be flexible towards various experimental setups and includes covariate features, such as ionization mode, a continuous scale of collision energies and compatibility with many types of MS instruments. The training and test datasets are aggregated from multiple sources and much effort has been put to accommodate a variety of MS experiments.

Data

FIORA is trained on a merged library from NIST (2017) and MS-Dial [34]. The latter is in itself a collection of various spectral libraries that provides accessible metadata, such as collision energies, that FIORA utilizes. Two 10% splits are taken for validation and testing, respectively. Furthermore, the spectral predictions are evaluated on three separate datasets that are disjoint from the set of training and validation compounds. The library test split serves as a distribution of "unknowns" from the same sources. In addition, spectra from the CASMI challenges in 2016 and 2022 provide a more independent distribution of unknown compounds. As part of the challenge, the experimental MS/MS data were used to test *in silico* algorithms in their ability to identify compounds that were not recorded in spectral libraries. Although spectral reference libraries have since been expanded, the CASMI challenge compounds remain important cornerstones for cross-referencing the performance of metabolomics software and have been used for benchmarking purposes in many studies [23, 31]. All compounds from the test sets were explicitly excluded from the training process. For further details refer to the Data preparation and Training and testing sections.

Model selection

FIORA shows versatility by not being constrained to a single model architecture. Its modular design allows for multiple prediction targets and effortless integration of different deep learning architectures, as will be evaluated in this section. Note that while FIORA learns all prediction targets, i.e., fragment intensities, RT, and CCS values, our primary focus is on the spectral predictions in terms of hyperparameter tuning and model selection.

The initial model selection is conducted on the validation split, examining the performance of popular graph network architectures with variable network depths. The effect of different architectures on cosine similarity (using square root intensities, as described in Evaluation metrics) between the spectral reconstruction and the ground truth validation spectra is shown in Figure 2. All other hyperparameters are fixed. Interestingly, the graph convolutional network (GCN) and relational graph convolutional network (RGCN) outperform attention-based networks, i.e., GATs and Transformers, especially as the network depth increases. There is a sweet spot in graph depth at 4 to 6 layers, which maximizes cosine similarity to the validation set. Notably, 0 or 1 graph layers are significantly less powerful because very little structure information is aggregated. Similarly, a high number of graph layers (>7) leads to reduced performance. GNNs are known to lose expressive power when too many graph convolutions are applied, as the hidden node representations become indistinguishable [35]. This is particularly evident for the attention-based mechanisms, suggesting that they may be less effective at predicting fragment ions. The integration of bond type information to graph convolutions, as seen in the RGCN compared to the GCN, appears to have a small positive impact on prediction quality for high network depth.

Note that fragment ion prediction is centered around bond breaks. This process incorporates bond (edge) features, node embeddings of the two neighboring atoms and covariate features, such as collision energy, at the final layers. This way, at depth 0 the predictor is aware of the bond type and the connected atoms. Similarly, at depth 5, substructure information of up to 6 atoms from either side is aggregated, thereby covering a complete 6-cycle ring structure. The RGCN with a depth of 6, which performed best on the

validation set, was selected for subsequent benchmarking on the three test sets. Exact model specifications can be found in the Methods section and the Supplementary Material.

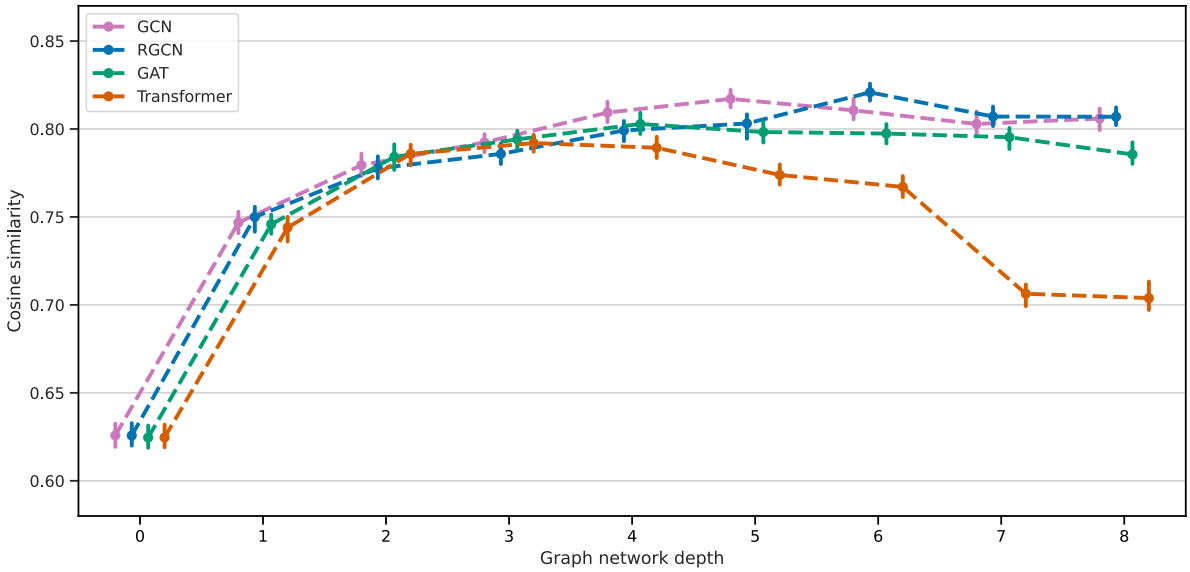

Figure 2: Median cosine similarity evaluated on the validation split. Error bars show the 95% confidence interval. Cosine similarity reaches its peak between a network depth of 3 and 6 layers before it falls off again. The graph convolution (GCN) and relational graph convolutional networks (RGCN) perform better than the attention based mechanisms (GAT and Transformer).

Spectral prediction quality

Of the algorithms discussed in the Related work section, only CFM-ID and ICEBERG are *in silico* fragmentation algorithms in a true sense, i.e., methods annotating fragment structure and modeling break events. Recently, Goldman *et al.* [31] conducted a comprehensive benchmarking study, in which ICEBERG outperformed all other spectral prediction software. For this reason, we compare FIORA with ICEBERG and CFM-ID.

Table 1 shows the overall cosine scores, separated for positive [M+H]+ and negative [M-H]- precursor charges. Other types of precursors or adducts are not supported by FIORA at the moment. Note that ICEBERG was retrained on the exact same dataset as we trained FIORA on (for positive spectra exclusively). For CFM-ID this was not feasible and the latest model, pre-trained on the METLIN library [17], was used instead.

FIORA's predicted MS/MS spectra exhibit the highest median cosine similarity to reference test spectra for the test split and CASMI 16 dataset, with a gain ranging from 10% to 44% over the runner-up. The relative improvements are more pronounced for the negative test sets. For the CASMI 22 dataset, the overall cosine scores are significantly lower for all algorithms compared to the other test sets. FIORA is slightly better than CFM-ID in the negative ionization mode, but falls short in the positive mode. The performance between ICEBERG and CFM-ID is similar, with ICEBERG being slightly better on the test split and CASMI 16, but CFM-ID being superior on the CASMI 22 dataset. Essentially, we could not observe a clear advantage of the newer method, ICEBERG, over CFM-ID. This may be due to differences in the training dataset. Keep in mind, that our filtered training library is smaller than the library ICEBERG was trained

Table 1: Median cosine similarity of spectral predictions to ground truth test spectra. The columns are arranged according to the test sets and precursor ion modes (positive and negative). ICEBERG operates in positive ionization mode only. For more information on the test sets, please refer to the Training and testing section.

	Test split +	Test split -	CASMI 16 +	CASMI 16 -	CASMI 22 +	CASMI 22 -
Unique compounds	895	437	381	139	160	98
FIORA	**0.84**	**0.82**	**0.78**	**0.79**	0.25	**0.30**
CFM-ID	0.67	0.57	0.70	0.59	**0.38**	0.29
ICEBERG	0.72	-	0.71	-	0.36	-

on [31] and different from the METLIN database [17] used by CFM-ID. Another explanation could be that CFM-ID predicts spectra at different collision energy levels, whereas ICEBERG essentially predicts average spectra. Therefore, CFM-ID can replicate experimental conditions more closely. FIORA takes this idea even further by using continuous collision energies as an input parameter, which might be another reason for FIORA's overall stellar performance.

Interestingly, the low cosine scores for CASMI 22 coincide with the results reported by Goldman *et al.* [31]. They suggest that the poor performance may be due to an out-of-distribution bias. In fact, we also find that compounds in the CASMI 22 dataset have lower structural similarity to our training sets, compared to the 2016 dataset. Moreover, we identified 15 compounds from the CASMI 22 challenge in the initial NIST and MS-Dial spectral libraries, which were subsequently removed for test/training separation. This overlap allows us to examine differences between the spectral measurements recorded in the spectral (training) libraries and the CASMI 22 test set. A considerable number of MS/MS challenge spectra are inconsistent with data from NIST and MS-Dial. For instance, 25% of matching spectra have a cosine similarity of less than 0.1. To clarify, the matching library spectra present a completely valid test set with the same compounds as CASMI 22, but have very little spectral similarity to the actual CASMI 22 spectra, despite similar experimental conditions. This suggests that the CASMI 22 test set cannot be considered canonical. With that said, we deliberately report the CASMI 22 results to explore the limits of the different implementations. A section in the Supplementary Material is specifically dedicated to explaining the discrepancies and intricacies with the CASMI 22 dataset, including a more nuanced performance analysis. It is evident that CASMI 22 contains examples that are difficult to model, predominantly spectra with little signal intensity that can be explained by single-step fragmentation. We identify this as the main reason FIORA's underperformance on this particular dataset, which we discuss thoroughly in The impact of single-step fragmentation section.

Overall, FIORA outperforms both ICEBERG and CFM-ID in all but one test set, and even surpasses CFM-ID on the "challenging" CASMI 22 dataset in negative ionization mode. The gain in cosine similarity is consistently higher for negative mode spectra, which could be attributed to training positive and negative spectra in parallel with the same model. In contrast, CFM-ID uses a separate model trained for positive and negative spectra.

FIORA generalizes well across compound classes and to unknown compounds

Fragmentation algorithms, unlike spectral prediction models, have the unique ability to explain their output by virtue of their algorithmic design. Specifically, predicted peaks are annotated by fragment ions (and potentially fragmentation pathways), grounding the predictions in a relatable and physics-inspired process. However, due to the complexity of deep learning models with millions of parameters, it is often unclear how models generate their output. It is important to determine whether models take shortcuts or overfit to specific input types or features. In this section, we provide an overview of FIORA's ability to generalize to unknown compounds, evaluate the performance across different compound classes, and contextualize the latent feature representation acquired by FIORA's graph module with the structural properties of the compounds under study.

To assess the extent to which the fragmentation methods can generalize to uncommon structures, we compute the Tanimoto similarity (Jaccard index) between all test and training compounds based on their Morgan fingerprints (2048 bits; radius 3). The maximum Tanimoto similarity for each test compound serves as a meaningful descriptor of its structural similarity to the entire training set. Compounds with low Tanimoto similarity are arguably more difficult to model and represent a distribution of unreferenced compounds that are dissimilar to those in the spectral libraries. These compounds are of particular interest for spectral prediction because they cannot be easily related to other reference compounds using methods such as Molecular Networking, and constitute the unexplored chemical space, i.e., metabolomic "dark matter" [4]. Figure 3 depicts the median cosine similarity at different levels of Tanimoto similarity. Interestingly, FIORA's prediction quality remains stable with a median cosine similarity of above 0.85 for Tanimoto similarities above 0.6 and decreases only at lower Tanimoto similarity levels. Even for the most dissimilar set of compounds, the median cosine similarity is above 0.7, indicating excellent performance when generalizing to unfamiliar structures. The curve depicting ICEBERG's performance is very similar to that of FIORA, but with an overall lower cosine similarity. ICEBERG appears to have difficulty predicting spectra for compounds with a very low Tanimoto similarity of 0.2 to 0.3. CFM-ID was pre-trained on a different dataset, so the intervals do not correctly reflect the Tanimoto similarity between training and test compounds. This is evident in Figure 3, where there is a lack of a clear upward trend for CFM-ID and wider confidence intervals (as seen in the error bars). However, CFM-ID takes on the role of a

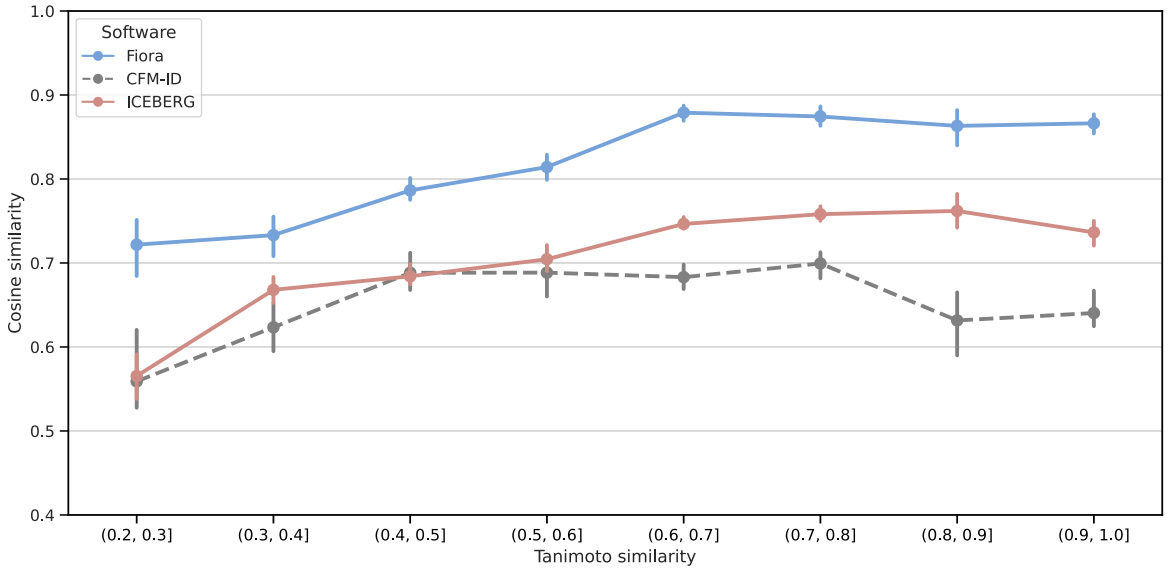

Figure 3: Cosine similarity at intervals of structural similarity of compounds from the test split to training compounds, measured by maximum Tanimoto similarity (Jaccard index) using Morgan fingerprints with 2048 bits and a radius of 3. Results are shown for positive ionization spectra to ensure the same dataset for all algorithms. Since CFM-ID was trained on a different dataset, the intervals do not reflect the actual Tanimoto similarity for the model. CFM-ID's cosine scores still provide an overview of the prediction performance for each interval, evaluated with a more independent model.

(more) independent evaluator of the different subsets of compounds. Importantly, the performance of CFM-ID is also lower for low Tanimoto similarities between 0.2 and 0.4, indicating that these compounds are either in fact rather uncommon or at the very least challenging to predict. We conclude that Fiora generalizes quite well to structurally dissimilar compounds, but shows a visible drop in quality for compounds with Tanimoto scores below 0.6, as expected. Still, the relative loss of performance loss is lower for Fiora (18%) than for ICEBERG (26%), when comparing the best performing interval to the interval with the lowest Tanimoto similarity. For the CASMI challenges this trend is even less pronounced (as shown in the Supplementary Material).

Moreover, the prediction quality of Fiora is very stable across different compound classes. Figure 4b shows the cosine similarity scores for individual compound superclasses, annotated using ClassyFire [36]. Fiora consistently achieves a median cosine score well above 0.7 for all compound superclasses, except for *organohalogen compounds*, which have a median score of 0.61. However, there are only three unique *organohalogen compounds* (11 spectra) in the test set, so this result carries little statistical weight. Similarly, *nucleosides, nucleotides and analogues* have an extremely high cosine similarity of 0.88 based on only 5 unique compounds (19 spectra). Overall, Fiora's performance appears to be robust across the test set without emphasizing specific compounds at the superclass level.

At the same time, the shared molecular structures within the compound superclasses have a significant impact on the latent representation that Fiora learns. Figure 4a shows Fiora's graph embeddings (mean pooled over all nodes) after a UMAP dimensionality reduction to 2-D [37], with each compound colored according to its corresponding superclass. Keep in mind that Fiora is not trained to produce a meaningful compound representation that can be used for property prediction, but rather to solve the edge break problem using local neighborhoods. Graph layers are not affected by the training of RT and CCS values, making the global molecular structure embedding purely a by-product of the fragmentation method. It is all the more impressive to see that this still results in compound embeddings that form structural clusters in Figure 4a, which can be broadly separated by their superclasses. The division goes beyond the superclass level. For instance, *Lipids and lipid-like molecules* are grouped together in two main clusters (seen at the bottom left and bottom right of the UMAP). Upon closer examination, one cluster is dominated by *Glycerophospholipids*, *Fatty Acyls*, and *Sphingolipids*, while the other cluster contains *Prenol lipids*, *Steroids*, and *Saccarolipids*. These lipid classes also separate well within each cluster, which is shown in the Supplementary Material.

Naturally, the clusters are also the result of similarities in the molecular graphs that Fiora receives as input. These may have similar element compositions or share certain structural elements within the superclasses. Nevertheless, it is important to recognize that

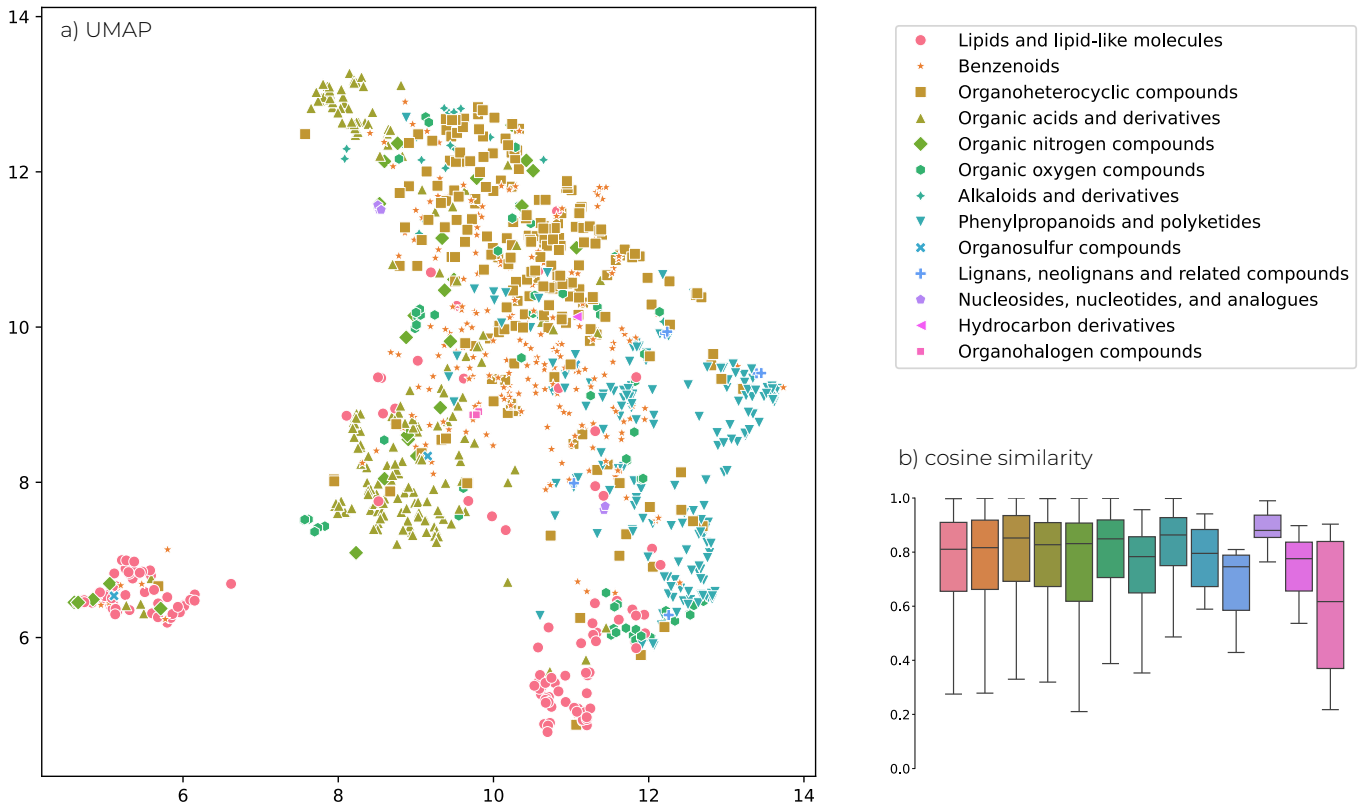

Figure 4: a) Uniform Manifold Approximation and Projection (UMAP) [37] visualization of the molecular graph embeddings. Each point corresponds to a unique compound and is color-coded based on compound superclasses, annotated by ClassyFire [36]. Clustering of compounds according to their superclasses is evident, indicating the learned structural relationships by FIORA during training. b) Cosine similarity between experimental and predicted test spectra split according to the compound superclasses. Prediction performance is consistent across most superclasses, with median scores of above 0.7.

the model calculates peak probabilities based on a structural foundation and even retains a global representation of the molecule that aligns with its class annotation. We also observe considerable improvements in the prediction of CCS values using this graph embedding, presented in the Retention time and collision cross section section, which indicate a high representational power of the embedding process. FIORA not only produces structurally meaningful embeddings, but also encapsulates critical information about the 3-D structure (CCS) and chromatographic properties (RT), and quite possibly other pertinent molecular properties as well.

Retention time and collision cross section

FIORA's architecture was designed to support additional prediction targets through individual submodules branching off the graph convolutional layers. As a proof of concept, we show that FIORA can accurately predict RT and CCS values. The model was trained on a small dataset of 409 compounds with RT information and 1346 compounds with CCS values from the MS-Dial library. It is important to consider the limited size of the training set when interpreting the results. Therefore, the performance was estimated conceptually and not benchmarked against state-of-the-art algorithms. It is worth noting that both RT and CCS values warrant dedicated studies for optimization and evaluation, as demonstrated in the study by Domingo-Almenara *et al.* (2019) on the Metlin small molecule retention time (SMRT) dataset [38].

Figure 5 presents parity plots for RT and CCS values, comparing FIORA's predictions to the experimental measurements. In terms of RT prediction, the majority of RT predictions fall within a 30-second deviation, although a non-negligible number do not. This indicates that the performance is somewhat inconsistent. The Pearson correlation coefficient (r) between the predictions and observations is 0.82 and an R^2 value is 0.65. Domingo-Almenara *et al.* report a good performance with a median absolute deviation of 35 seconds on the SMRT dataset [38]. However, the datasets are not necessarily comparable. All RT values in our study come from the BMDMS-NP library [39], which is a part of the MS-Dial spectral library. The exact experimental setup and gradients used for chromatographic separation remain unclear. The results suggest that RT prediction with FIORA is possible, but requires extensive retraining with a larger

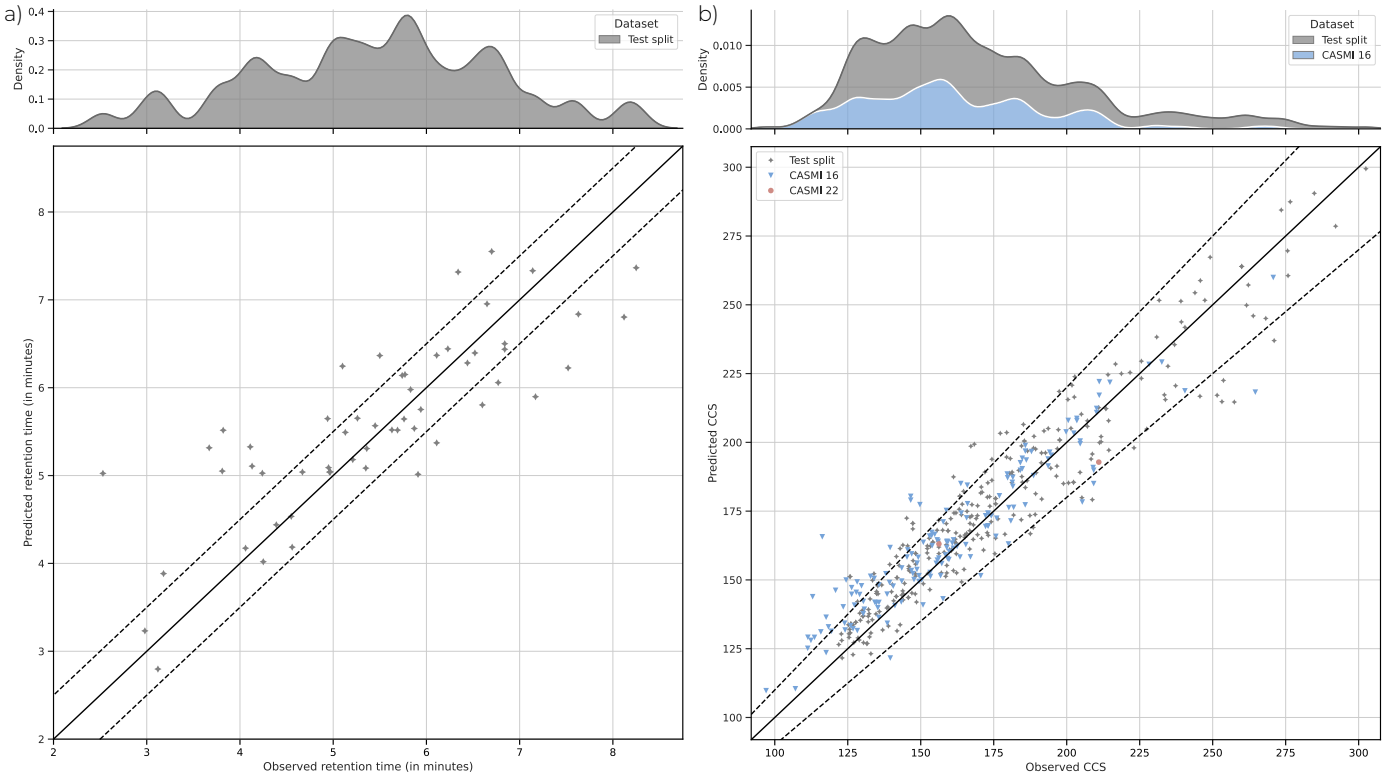

Figure 5: Parity plot of RT a) and CCS b) predictions by FIORA. Retention time values for the test sets were retrieved from the BMDMS-NP library [39] and CCS values from the whole MS-Dial library [34]. The diagonal lines describe perfect prediction. The dashed lines indicate a 30 second deviation for RT and a 10% deviation for CCS values from the ground truth observations.

and more homogeneous dataset.

Predicted CCS values are shown for all three test sets in Figure 5b. CCS values for CASMI 16 and CASMI 22 compounds could be partially annotated using the MS-DIAL library as reference, although only 2 compounds were found for CASMI 22. The vast majority of predictions fall within a 10% error range. To validate the performance, we compared the predicted CCS values with those estimated by a linear regression model based on ion mass. This comparison is important because molecular weight is a logical proxy of CCS. FIORA achieves a very high Pearson correlation coefficient (r) of 0.97 and R^2 value of 0.95 for the test split, which is slightly better than linear regression (with r= 0.95; R^2=0.9). Notably, for CASMI 16 compounds, FIORA predictions are significantly better than linear regression with a Pearson correlation coefficient of 0.96 and R^2 of 0.92 (compared to r=0.79 and R^2=0.61). Thus, FIORA appears to generalize much better to a more independent distribution of test compounds. It remains to be seen whether the additional orthogonal MS features incorporated by FIORA translate to improved compound identification rates.

Significant speed improvements through GPU acceleration

Run time is a particular weakness of fragmentation algorithms, primarily due to the combinatorial nature of potential bond breaks. At the same time, fast processing time is critical to cover the vast chemical space of known structures with spectral predictions and for high-speed (re)scoring pipelines of putative candidates. Table 2 shows the total and average prediction time measured for the algorithms across all test sets. FIORA runs approximately 20 times faster on the GPU than on the CPU, and predicts around 10,000 spectra within just five minutes on an NVIDIA A100 GPU. On CPU, FIORA is still 4.6 times faster than CFM-ID but slightly slower than ICEBERG.

Compared to the run times reported by Goldman *et al.* [31], we observe a slightly lower average prediction time for ICEBERG and a significantly lower prediction time for CFM-ID. Note that CFM-ID always predicts three energy levels per compound. This was taken into account when calculating the number of predicted spectra, but also makes the average run time a more generous metric for CFM-ID. FIORA stands out as the only software capable of predicting spectra for all compounds at all collision energies. Note that FIORA could be further optimized for run time, e.g., by adding mini-batches to the prediction process. This was realized for the training loop,

Table 2: Run time comparison over all test sets. The number of predictions vary due to the specifications of each algorithm. FIORA is the only software that predicted all compounds at all collision energies.

	Total time	Spectra predicted	Average time per prediction
FIORA (GPU)	**4m 58s**	9725	**0.03s**
FIORA (CPU)	1h 40m 38s	9725	0.62s
CFM-ID	4h 24m 29s	5544	2.86s
ICEBERG	46m 34s	6146	0.45s

where 200 epochs of training and validation (without early stopping) were completed in 3 hours. Additional training of RT and CCS values took under 10 minutes. In comparison, ICEBERG was trained on only the positive spectra on a GPU for over 6 days. All in all, by taking advantage of GPUs, FIORA outperforms the other methods by a wide margin in terms of training and prediction speed.

The impact of single-step fragmentation

Despite the performance benefits of FIORA, its apparent biggest limitation is the shift towards single fragmentation. This section analyzes the impact of this decision on prediction quality. It should be noted that FIORA implicitly covers multiple bond breaks from the same residue, as is explained in the Fragmentation algorithm section. Still, the model does not account for fragments that break off from different sides of the molecules or cleavages of circular structures. As a result, FIORA does not cover a fragment space as comprehensive as that of CFM-ID or ICEBERG, which may ultimately lead to a significant amount of missed signal for some compounds. Figure 6 displays the impact of peak intensity coverage on the cosine similarity of the predicted spectra. We define peak intensity coverage as the fraction of peak intensity that can be theoretically explained by FIORA's fragmentation algorithm, rather than the actual amount of peak intensity matched by the prediction, although the two are very related as long as the fragment prediction is accurate. Therefore, coverage describes the relative peak intensity covered by a perfect prediction. The maximum cosine similarity that can be reached is bounded by the square root of the coverage, shown by the dotted line in Figure 6. Note that this is an optimistic upper limit, which is slightly lower when peak intensities are distributed more uniformly among multiple peaks. Hence, the individual maximum for each compound spectrum might be even lower than indicated. It is remarkable how accurately FIORA predicts the intensities of those peaks that can be matched. A significant fraction of the predictions have a cosine score close to the maximum. In other words, a loss in predictive performance can often be attributed to reaching the theoretical optimum rather than poor peak intensity prediction. Despite this limitation, FIORA performs exceptionally well compared to state-of-the-art methods (refer to Spectral prediction quality). We have also shown in the section FIORA generalizes well across compound classes and to unknown compounds that our approach does not lead to major performance differences between compound superclasses or for structurally distinct compounds.

However, low coverage still has a noticeable effect on the overall performance. This is particularly evident in Figure 6 (top panel) for the CASMI 22 dataset, where most compound spectra have an intensity coverage of close to 0. This is the primary reason for FIORA's weak performance on this dataset. Higher collision energies exacerbate the problem, as shown in the Supplementary Material. Explicitly modeling continuous collision energies is a key feature FIORA has over the other algorithms. Nonetheless, it is strictly limited to peak intensities of the same set of fragments and becomes ineffective when too many fragmentation events occur. Remember that we have already pointed out inconsistencies in the CASMI 22 data in the Spectral prediction quality section, so the low coverage is likely influenced by an abundance of noise peaks or poor spectral quality. This theory is supported by the low cosine scores of CFM-ID and ICEBERG, despite multi-step fragmentation, and by the additional data provided in the Supplementary Material. As such, this is the exception rather than the norm. High coverage, as seen in CASMI 16 and the test split, is directly correlated with a significantly higher prediction quality for FIORA.

While FIORA undoubtedly faces a limitation with its fragmentation method, it effectively compensates by leveraging graph substructures and covariate information to near perfection, resulting in very accurate intensity predictions. Compared to the current state of the art, FIORA performs exceptionally well. At the same time, it should be noted that single-step fragmentation will always constrain FIORA for certain compounds, and it is an important milestone for future improvements.

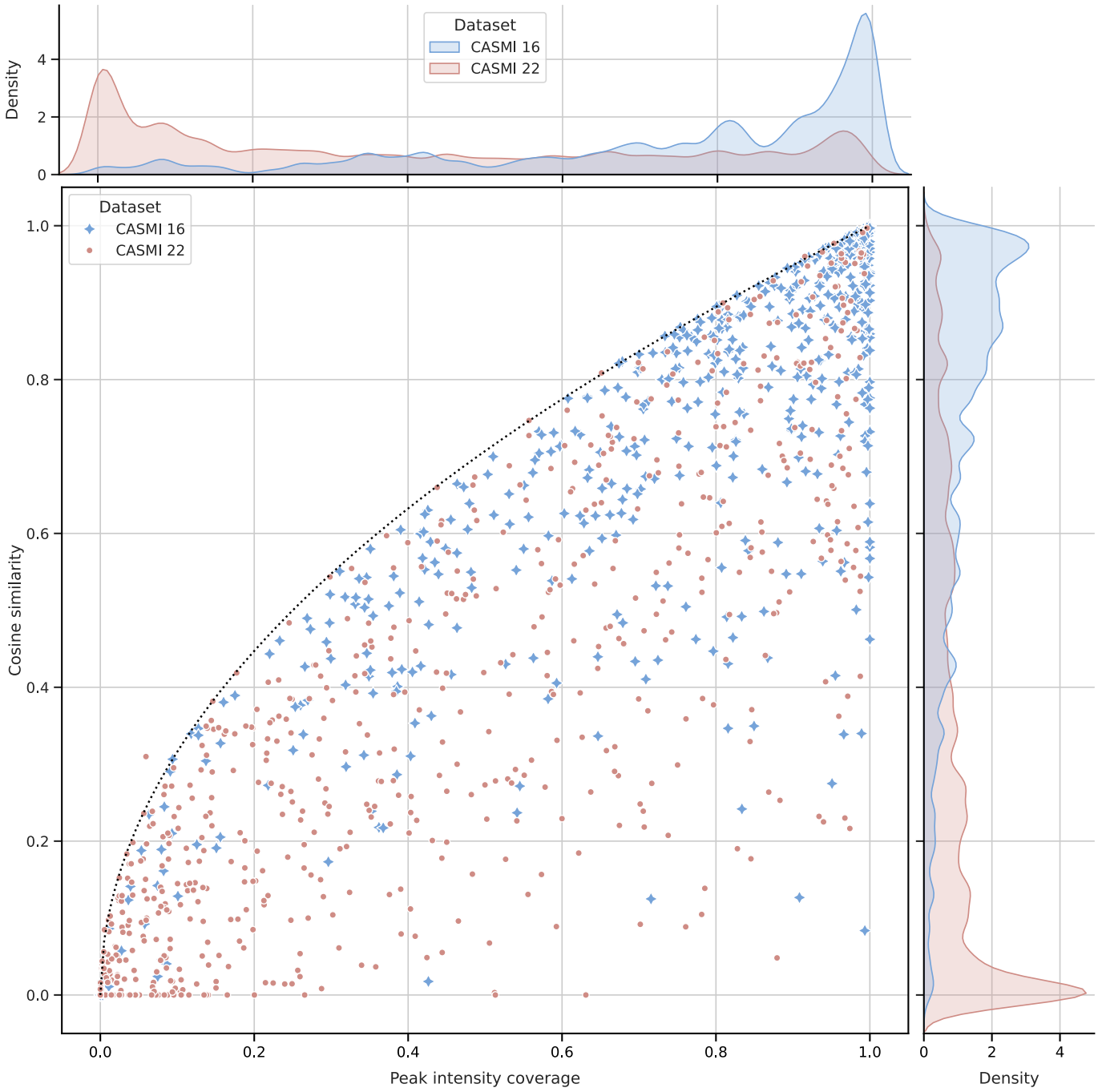

Figure 6: Cosine similarity over peak intensity coverage for predicted spectra from the CASMI 2016 and 2022 challenges. The dotted line describes an optimistic upper bound, i.e., maximum cosine similarity at that specific intensity coverage. Occasional outliers above arise from differences in the relative tolerance (50 ppm) FIORA uses for fragment annotation and the absolute tolerance (0.05 Da) used to calculate cosine similarity. CASMI 2016 represents a standard dataset with high coverage distribution, resulting in overwhelmingly high cosine scores. CASMI 2022 represents a rare low-coverage scenario, which leads to low cosine scores.

DISCUSSION

Experimental spectral libraries are always going to be incomplete in some way [15]. *In silico* generated spectra can complement these libraries. The nature of fragmentation algorithms makes them inherently valuable for metabolomics beyond spectral prediction alone. Unlike spectral predictors that function as "black box" neural networks, fragmentation algorithms can utilize our understanding of the underlying physical processes and potentially expand our knowledge by learning from experimental data. They can aid compound identification algorithms by anticipating the fragment ion distribution and providing an orthogonal reference to evaluate compound

candidates. Bond breaks and fragment ions can be judged individually, allowing for manual and systematic validation of compound annotation. By mirroring the physical fragmentation process, the algorithms lend credibility to the simulated mass spectra, which is critically needed when exploring new compounds that are not included in reference libraries.

In this work, we introduce FIORA, a novel fragmentation algorithm that advances the field in a number of key ways. We show that fragment intensity predictions are significantly improved by modeling bond dissociations based on their local molecular neighborhood. Despite having a smaller set of fragments compared to the state-of-the-art algorithms, FIORA yields higher cosine similarity scores across the majority of test datasets. Notably, already a small number of graph convolutions, describing short-range structural relationships, produce models with high predictive power (refer to Figure 2). This achievement should be taken into consideration for all future implementations of *in silico* fragment ion predictions. We would like to emphasize that our unique approach of learning local substructures surrounding the bonds rivals global molecular embedding strategies, which are commonly used in recent spectral prediction models, including ICEBERG. Additionally, FIORA incorporates covariates at the fragment intensity prediction level, including ionization mode, instrument type, molecular weight, and collision energy. Collision energy, in particular, has a significant impact on peak intensities as an increase causes the precursor to diminish and new, smaller fragment ions to emerge in the mass spectrum. For example, when modeling the CASMI 2016 data, merging the spectra predicted at the three collision energy steps used in the experiment results in a more accurate simulation than simply predicting the average collision energy (as shown in the Supplementary Material). In fact, predicting a spectrum at any given collision energy is already a significant improvement over the fixed energy levels provided by CFM-ID.

Combining the training of positive and negative spectra into a single model allows the algorithm to learn from the fragmentation patterns of the other ionization type. As a result, the performance gain over CFM-ID is even greater for negative spectra (refer to Table 1), despite having twice as many positive spectra to learn from as negative ones. Overall, FIORA was trained on a relatively small dataset of approximately 10,000 unique compounds, greatly limited by the lack of experimental metadata, especially with respect to the collision energies applied. This might be expanded in the future with better standardization of workflows and ever-growing spectral libraries. Despite the more focused dataset, FIORA generalizes well across different compound classes and to structurally distinct compounds (illustrated by Figure 3 and Figure 4). We show that FIORA is capable of learning meaningful representations and compound defining properties with its molecular graph embedding, which is generated as a by-product of the fragmentation process. Based on this embedding, molecules can be clustered at the compound class and superclass level. Simultaneously, the molecular embedding serves as an excellent starting point for the prediction of other MS-related attributes. We conceptually prove this through the prediction of RT and CCS values (shown in Figure 5). While this feature is still in its prototype stage, it should be considered for the future of spectral prediction software, as it provides valuable dimensions that can help distinguish candidate compounds. Whether it will actually improve compound identification workflows remains to be shown in follow-up studies.

Importantly, FIORA's algorithm maintains a remarkable level of interpretability at the molecular level, while the individual fragment ion predictions are comprehensible in their own right, as they are governed by only a small number of surrounding atoms. This level of explainability is a quality that is often sacrificed in the era of deep learning. In terms of run time, FIORA outperforms CFM-ID and ICEBERG considerably by taking advantage of GPU-accelerated computations (refer to Table 2) and can be further optimized in the future. Prediction speed is crucial due to the vast space of chemical species, which is poorly covered by spectral libraries.

With that said, no single algorithm is objectively superior to the others in every aspect. On the contrary, this study highlights the advantages and differences between the approaches and closely examines the input, methods, and various aspects of prediction quality. ICEBERG is a fast and effective algorithm that could benefit from integrating collision energies and additional features representing bond dissociation. CFM-ID is based on an 11-year-old algorithm, but remains relevant through a solid statistical foundation and consistent updates. Lastly, FIORA presents a fresh take on bond dissociation, but is limited by single-step fragmentation. In this way, FIORA 's fragmentation algorithm could be seen as less effective, since FIORA covers a smaller set of fragments. Indeed, the lack of multi-step fragmentation is currently the biggest limitation. In every other way, FIORA succeeds in setting a new state of the art in terms of intensity prediction quality and the integration of relevant molecular substructures. We would like to point out that our implementation leaves open a future extension to multi-fragmentation. The predicted fragments are graphs themselves and can be

recursively fed to the model for further fragmentation. However, it is difficult to link branching fragmentation pathways directly to the observed product ions, i.e., peaks, which serve as ground truth. More sophisticated statistical methods, as for example implemented in CFM-ID, would be required.

In conclusion, advances in machine learning and the ever-growing spectral libraries introduce a new era of fragmentation algorithms. Simulated MS/MS spectra may soon match the quality of experimental libraries and are critically needed to cover the large space of unreferenced chemical species. With this work, we make a pivotal contribution to the field of *in silico* fragmentation as FIORA taps into the full potential of molecular substructures.

METHODS

Fragmentation algorithm

FIORA is designed to take advantage of the unique power of graph neural networks to learn structural patterns and local neighborhoods around chemical bonds. Each molecular structure M is represented as a graph G with atoms for nodes and bonds for edges, which is common practise in computational chemistry. The molecular structure graphs are built from string representations, e.g., SMILES. FIORA operates on neutral molecular structures and only considers information about precursor charge ([M+H]+, [M-H]-) and other covariates at the very end. Fragment ions are modeled by the removal of edges in the graph, indicating singular bond cleavages.

A key concept of our method is that we explicitly model ion rearrangements through hydrogen losses. This is important for direct assignment of peaks to fragment ions and allows end-to-end prediction from the molecular structure graph G to the fragment ion space $\mathbb{F}(G)$. The latter is constructed as follows: Let $E(G)$ denote the set of edges in G and G_{-e} the pair of subgraphs (fragments) that arise from removing edge $e \in E(G)$. The fragment ion space is the set of all subgraphs and fragment ionizations, accounting for up to 4 hydrogen losses, i.e.,

$$\mathbb{F}(G) = \bigcup_{e \in E(G)} \bigcup_{F \in G_{-e}} \left\{ [F+H]^+, [F]^+, [F-H]^+, [F-2H]^+, [F-3H]^+ \right\} . \tag{1}$$

For each molecular graph G, FIORA predicts the precursor stability σ and the abundance values θ_f for all $f \in \mathbb{F}(G)$. Both are combined using a *softmax* function to compute *fragment probabilities*:

$$p(f) = \frac{\exp \theta_f}{\exp \sigma + \sum_f \exp \theta_f} . \tag{2}$$

Predicting abundance values θ_f is conceptually related to the idea of *break tendency* values that were proposed by Allen *et al.* 2015 [22], but here we extend this concept to individual fragment ions f. In addition, the precursor stability σ allows to model abundances of the intact molecule under various conditions, such as different collision energies. Negative precursor molecules are treated analogously with negative fragment charges.

The MS/MS spectrum is reconstructed afterwards from the exact fragment ion m/z. Figure 7 illustrates the fragment ion prediction process of a single edge break and depicts the information flow in the graph network.

It is important to note that FIORA can be very well used recursively, since fragments are graphs themselves. However, multiple fragmentation steps make it significantly harder to assign ground truth ion probability during training or require other stochastic modelling approaches, such as a Monte Carlo Tree Search. Every assignment of fragment probability (derived from training spectra) bares the risk of introducing conflicts. These arise, when more than one structure can explain a peak, either by having the same or similar weight within the mass tolerance. With multiple fragmentation pathways leading to the same fragment (or fragments of similar weight), the number of conflicts increases dramatically. In favor of speed and more harmonious ion probability assignments we reduce the possible fragmentation events to one. Nevertheless, multi-step fragmentation should be investigated to refine the model in the future. Importantly, multiple bond cleavages from the same residue are already implicitly modelled by directly assigning probabilities to the end-product ions (observed peaks), thereby keeping the missing signal low for the most part. The impact of missing peaks is discussed in The impact of single-step fragmentation section.

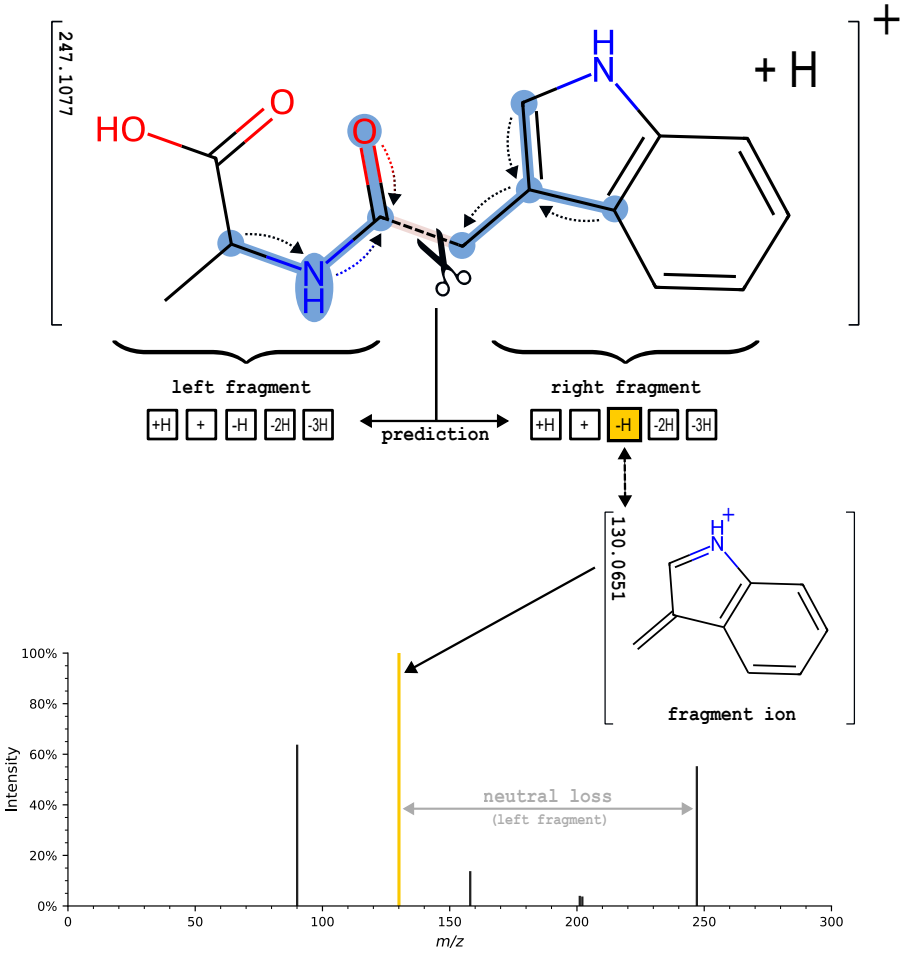

Figure 7: Depiction of FIORA's fragmentation algorithm exemplified for the central bond. FIORA learns the local neighborhood of the bond over several graph convolutions, as illustrated by the dotted arrows indicating the information flow. The learned molecular structure is outlined in blue. For visual clarity, only two graph convolutions are illustrated and arrows are directed towards the designated bond. Based on the bond features and surrounding substructure, fragment ion abundances (and neutral losses) are predicted. In this case, the prediction suggests a loss of two hydrogen atoms and the formation of a new double bond in the right fragment. Peak probabilities are reconstructed statistically considering precursor stability and the abundances of all other fragments.

Model architecture

FIORA uses a graph network model to predict the precursor stability σ and the fragment abundances θ_f from a molecular graph G. Since the fragment space $\mathbb{F}(G)$ is produced from individual edge breaks, edge prediction can be directly used to infer fragment ion abundance θ_f. The prediction of σ is based on the whole graph representation.

Initially, atom features are encoded as vectors of integer numbers (for element type, number of hydrogen atoms bound, and ring type information) and are passed to an embedding layer for each number. The embedding dimension of 300 was determined empirically. Bond features are embedded similarly based on bond type and ring type information.

Mathematically, a graph network layer involves a permutation equivariant function Q updating the node (atom) features X into a latent representation $H = Q(X, A)$ using the connectivity of the graph or adjacency matrix A. Definitions follow Bronstein *et al.* (2021) [40]. This is achieved by applying a shared permutation invariant function ϕ to all $x_i \in X$ and to the features of their local neighborhoods N_i, where ϕ is a learnable function. In the general case of a message passing neural network, the layer-wise update function can be formulated as

$$h_i = \phi(x_i, \bigoplus_{j \in N_i} \psi(x_i, x_j)), \tag{3}$$

where $h_i \in H$ describes the new hidden representation of node i, $\bigoplus$ is a permutation invariant aggregation operator, e.g., the sum ($\sum$), and ψ is a learnable *message* function. Based on the choice of ϕ and ψ, various types of graph layers can be modeled.

The type of graph network FIORA uses is customizable. Our implementation is based on the PyTorch Geometric library [41], which provides many graph network architectures. Currently, FIORA supports GCN [42], RGCN [43], GAT [44] and Graph Transformer layers [45]. After each layer an Exponential Linear Unit (ELU) activation function is applied. The choice of graph network and layer depth is discussed in Model selection. We use a 6-layer RGCN as default.

After applying the graph convolutions, the hidden node (atom) representations are taken as input to the final prediction using blocks of fully connected neural networks, which are separate for different prediction targets. For fragment ion prediction, we implement an edge map to concatenate and stack features of all the two node combinations connecting an edge. In addition, the respective edge embedding and covariates (molecular weight, precursor ion mode, collision energy, and instrument type) are concatenated. A fully connected neural network with two layers projects the input onto a 10-dimensional output vector of logits, representing fragment ion abundance θ_f for the 5 ion modes and for both sides of the modeled edge break. The precursor abundance σ is estimated from the entire graph embedding using a global mean pooling aggregation, concatenated with the same covariates. Two additional fully connected layers produce a single logit σ. All logits (abundance values) are concatenated and passed through a *softmax* function to model precursor and fragment ions as a probability distribution. For model training, a weighted mean squared error (MSE) loss is computed between the predicted and observed fragment probabilities. The latter are estimated in advance by matching peaks to the fragment space. The loss is weighted by 1 over the number of spectra available per compound to reduce the bias towards compounds with multiple entries in the libraries. Network parameters are optimized using ADAM [46].

Finally, the MS/MS spectrum is produced by tracing back predicted subgraphs (fragments) using the edge map and reconstructing the exact peak m/z from fragment weight and hydrogen losses. Ion probabilities are summed to obtain peak intensities because multiple ions can produce the same peak.

RT and CCS values are predicted based on the mean-pooled graph embedding. This is similar to precursor intensity prediction, but with a distinct set of weights and the standard MSE loss function. Hyperparameters were not specifically tuned for this task and training was performed after the fragment ion prediction, freezing all but the relevant dense layers that produce the RT and CCS estimates.

Data preparation

Experimental MS/MS data of metabolites is diverse. Different data formats and units must be aligned across all measurements, particularly with regards to metadata. Metadata describes experimental conditions and is crucial to understanding and predicting compound fragmentation. We made significant efforts to collect data from various sources and align the information as accurately as possible. This enables our model to utilize covariate information, such as collision energy or instrument type, but requires extensive data pre-processing. Nevertheless, data preparation is a critical factor influencing model performance. The following section provides detailed information on the most important pre-processing steps.

We use two libraries in this study. The first is NIST17, which contains a large collection of MS/MS spectra acquired from authentic compound standards. The library was converted to .msp format using the *lib2nist* program that is included with the NIST library. Compound information was exported in the .MOL format and then parsed back into the spectral library using the python *rdkit* package. For a detailed walkthrough of all library parsing steps, please refer to our script on GitHub (found at `https://github.com/BAMeScience/fiora/blob/main/lib_loader/nist_library_loader.ipynb`). It is important to note that NIST17 is a commercial library requiring a license, which severely limits its public use. The homogeneity of the data may also limit its ability to generalize to more diverse experimental setups. The NIST17 MS/MS library does not provide information about RT and CCS values of the measured compounds.

In addition to NIST17, we process a public library available in .msp format on the MS-DIAL website [34], which contains a collection of annotated MS/MS spectra from various other spectral libraries, such as MassBank and GNPS. Unlike most of the public libraries, the MS-DIAL spectral library standardizes metadata, making crucial information about collision energy, RT and CCS readily accessible.

Both libraries are pre-processed in a similar fashion, converting normalized collision energies (NCE) to electron Volts (eV) where pos-

Table 3: Overview of the two spectral libraries used for training. Sources list only the biggest contributions. Additional information is found on the provider websites `https://www.sisweb.com/software/ms/nist17.pdf` and `http://prime.psc.riken.jp/compms/msdial/main.html`.

	Library	**NIST 17**	**MS-Dial**
	Publicly available	No	Yes
	Sources	NIST	Massbank, RIKEN, GNPS, MoNA, BMDMS-NP ...
	RT information	No	Partially
	CCS values	No	Partially
before filtering	Entries	574,826	368,860
after filtering	MS/MS spectra	54,814	19,587
	Unique compounds	7,271	4,408

sible. MS/MS spectra with missing collision energies or unclear formats were excluded. Compounds with inconsistencies between SMILES, MOL format and InChiKeys were filtered out immediately. In addition, we filtered for ionization types [M+H]+ and [M-H]- and imposed requirements for spectra to be selected for training. These requirements included a maximum weight of 1000 Da and maximum collision energy of 100 eV. Compounds were fragmented by single edge removal, and peaks were matched at 50 ppm to set of fragment ions described in the Fragmentation algorithm section. The fragment matches are then used to determine the final data set containing spectra suitable for training. We enforced at least two peak matches, a minimum peak intensity coverage of 50% and at most 90% precursor intensity. Then, soft filters required that each spectrum have either 50% of the total peaks or at least five peaks matched by fragments, or an exceptionally high intensity coverage of 0.8. This ensured that there is sufficient overlap between the theoretical fragmentation patterns and the observed peaks. All this results in a combined dataset of 74,401 spectra covering 10,692 compounds. An overview of the libraries can be found in Table 3.

Training and testing

FIORA was trained on 80% of the molecular structures, with the remaining 20% split evenly between validation and testing. The training process consisted of 200 epochs using the ADAM optimizer, a weighted MSE loss and a scheduler that reduces the learning rate upon reaching a plateau in validation loss. The model checkpoint with the lowest validation loss was loaded afterwards. Hyperparameters were tuned mostly empirically using the validation set. The types of graph layer and model depth are systematically evaluated in the Model selection section. Exact model specification can be found in the Supplementary Material.

ICEBERG was trained using identical training and validation splits, and following the training steps provided at the original GitHub repository (`https://github.com/samgoldman97/ms-pred`). We used the commit from October 21, 2023, which is slightly newer than the 1.0.0 release, as it contains a detailed retraining workflow. In the case of CFM-ID, we used the pre-trained model v4.4.7 running with a docker container provided on Docker Hub (`https://hub.docker.com/r/wishartlab/cfmid`). Retraining CFM-ID with our dataset was found computationally infeasible.

In addition to the 10% test split from the NIST17/MS-DIAL library, we selected the CASMI 2016 and CASMI 2022 datasets for benchmarking. Both datasets were downloaded from the CASMI contest webpage at `http://casmi-contest.org`. We selected [M+H]+ and [M-H]- precursors and set covariates according to the descriptions. In each case, as many compounds and spectra as possible were extracted, which includes the priority and bonus challenge for CASMI 2022 and training as well as challenge spectra for CASMI 2016. In the CASMI 2016 challenge, a stepped collision energy was used, so we predicted spectra with FIORA for all 3 collision energies and merged them into a single "stepped" spectrum. CFM-ID predicts compound spectra at fixed collision energies of 10 eV, 20 eV and 40 eV. Therefore, we evaluated the prediction with the closest matching collision energy, following the approach of Wang *et al.* (2021) [23]. For CASMI 2022, the extraction of the challenge spectra was more complicated. The OpenMS [47] library using its Python wrapper was used for the reading the *mzml* file and for the extraction of the collision energy levels. Challenge spectra were extracted using a precursor tolerance of 10 ppm and retention time window of 5 seconds. Multiple measurements at the same collison energy were merged into a consensus spectrum. Despite our efforts, not all collision energies for all challenge spectra were found. Reducing the stringency of the tolerance values did not solve this, as it lead to ambiguous matches to more than one compound. All processing steps are found on our GitHub (`https://github.com/BAMeScience/fiora/tree/main/lib_loader`). To preserve the integrity of the test results, all compounds from the test sets were removed from the training data for Fiora and ICEBERG. CFM-ID was trained on the METLIN [17] library, which means that this separation is not guaranteed.

Evaluation metrics

Cosine similarity (Equation 4) measures the similarity of two vectors, A and B, in the inner product space. It is defined as the cosine of the angle ρ between the two vectors:

$$\text{Cosine similarity}(A, B) = \cos(\rho) = \frac{A \cdot B}{\|A\|\|B\|}. \tag{4}$$

To obtain the cosine similarity between MS/MS spectra, they need to be discretized along the m/z dimension. This is typically done by binning m/z values or by matching query and reference peaks and then assigning a dimension to each peak pair and each unmatched peak. The latter is specified in Equation 5, which defines the cosine similarity between the spectra S_A and S_B at a tolerance value t that determines whether two peaks match:

$$\text{Spectral cosine similarity}(S_A, S_B, t) = \frac{\sum_{(mz_k, I_k) \in S_A} \sum_{(mz_l, I_l) \in S_B} I_k I_l \mathbb{1}_{|mz_k - m_l| \leq t}}{\sqrt{\sum_{(mz_k, I_k) \in S_A} I_k^2} \sqrt{\sum_{(mz_l, I_l) \in S_B} I_l^2}}. \tag{5}$$

Peaks are represented as tuples of m/z and intensity values (mz_k, I_k). Note that Equation 5 reflects the cosine angle only if peak matches are unique. This is not necessarily the case as multiple fragment predictions may have the same or very similar m/z. Therefore, intensities are summed within the tolerance window defined by t, only for the purpose of spectral scoring.

Although the cosine similarity expertly describes the angle ρ between the spectral vectors in an inner product space, it has some practically shortcomings. Dominant (high-intensity) peaks have a dramatic effect on the cosine similarity, regardless of whether other low-intensity peaks are matched. As a result, the vanilla cosine similarity may not be indicative of how well low-intensity fragmentation patterns are reflected. This problem is exacerbated by a small number of peaks. Solutions are hyperscores, which take into account the number of matched peaks and are used in proteomics [48] or logarithmic or square root transformations of the initial peak intensities, which even out differences in peak intensity and de-emphasize dominant peaks. In this study, all results use the cosine similarity of square root transformed peaks, as is common in metabolomics.

However, whether common practise or not, it is imperative to understand that such a decision is arbitrary and that other similarity metrics may work equally well or better for the purpose of compound identification. In a previous study, we showed that the standard cosine similarity (with square root intensities) for predicted spectra is deceptive in some cases and leads to high scores for false spectrum matches [49], albeit for peptide spectra. In metabolomics, many of the same principles still apply when assessing spectral similarity for small compounds. A metric such as the cosine bias, described for spectra by Lam *et al.* (2007) [50], is a good indicator of how dependent the cosine similarity is on matching a few dominant peaks. Ideally, a high quality prediction would have a high cosine similarity and a low bias due to many matching peaks. We have suggested a bias-adjusted cosine score before. However, many compounds spectra have only a few peaks and even a perfect match might result in a high cosine bias, making it extremely challenging to fine-tune scoring functions. We have monitored the bias for different versions of the cosine similarity in the Supplementary Material.

Other adjustments to the cosine score include removing the precursor, reweighting peaks based on ion mass, and removing of unmatched query peaks. The former may eliminate a dominant precursor peak in some cases, but significantly increases the overall cosine bias. Second-order dominant peaks may arise from obvious losses, such as H_2O, and the molecular stability and the relationship of the intact molecule to the fragments are a indispensable parts of the fragmentation pattern. This makes precursor removal rather futile for spectra that already contain very few peaks. Placing additional emphasis on higher ion masses, as suggested by Stein & Scott (1994) [51], for instance, by multiplying peak intensities by m/z values, makes sense in order to increase the score for more compound-specific fragments. Fragments ions with low masses are more likely to be shared between different compounds. However, we observe a slight increase in the cosine bias, and since none of the models are optimized to specifically favor heavier fragments, it is not the optimal metric to assess spectral prediction quality. In contrast, removing unmatched query peaks may be an effective method to estimate the accuracy of peak intensity prediction, considering only the fragment ion space generated by the algorithm. However, since FIORA's fragment space is smaller than that of the other tools, this might give an unfair advantage to FIORA. Any combination of scores or biases is worth further investigation.

In summary, as datasets and search spaces grow rapidly, refined similarity scores will be necessary to distinguish true spectral matches from false ones. This is especially true, since prediction software might not cover the complete fragment space of experimental spectra. At this point, we evaluate prediction quality solely on the standard cosine similarity using square root transformed intensities.

ACKNOWLEDGMENTS

The authors thank Justin van der Hooft, Niek de Jonge, Tanja Holstein, and Sasan Amariamir for all their helpful and stimulating discussions. Additionally, they sincerely acknowledge Ferdous Nasri for proofreading the manuscript.

AUTHOR CONTRIBUTIONS

Y.N., P.B., and T.M. conceived the initial idea for the method. A.K. assisted with the mathematical conceptualization. The analytical methods and results were verified by F.R. and J.L. with particular focus on the chemical and metabolomic aspects. Y.N. implemented the algorithm, wrote the manuscript, and created the visualizations. All authors discussed the results and contributed to the final manuscript.

AUTHOR COMPETING INTERESTS

The author declare no competing interests.

REFERENCES

1. Nash, W. J. & Dunn, W. B. From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC Trends in Analytical Chemistry* **120,** 115324 (2019).
2. Glish, G. L. & Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nature reviews drug discovery* **2,** 140–150 (2003).
3. Ho, C. S. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical Biochemist Reviews* **24,** 3 (2003).
4. Da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* **112,** 12549–12550 (2015).
5. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proceedings of the National Academy of Sciences* **112,** 12580–12585 (2015).
6. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods* **16,** 299–302 (2019).
7. Tsugawa, H. *et al.* Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical chemistry* **88,** 7946–7958 (2016).
8. Lai, Z. *et al.* Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature methods* **15,** 53–56 (2018).
9. Van Der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences* **113,** 13738–13743 (2016).
10. Schymanski, E. L. *et al.* Critical assessment of small molecule identification 2016: automated methods. *Journal of cheminformatics* **9,** 1–21 (2017).
11. Lab, F. *CASMI 2022 - Results* https://fiehnlab.ucdavis.edu/casmi/casmi-2022-results (2023).
12. Tanaka, W. & Arita, M. Physicochemical Prediction of Metabolite Fragmentation in Tandem Mass Spectrometry. *Mass Spectrometry* **7,** A0066–A0066 (2018).
13. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **47,** D1102–D1109 (2019).
14. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research* **46,** D608–D617 (2018).
15. Ludwig, M., Fleischauer, M., Dührkop, K., Hoffmann, M. A. & Böcker, S. De novo molecular formula annotation and structure elucidation using SIRIUS 4. *Computational Methods and Data Analysis for Metabolomics,* 185–207 (2020).
16. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature biotechnology* **34,** 828–837 (2016).
17. Guijas, C. *et al.* METLIN: a technology platform for identifying knowns and unknowns. *Analytical chemistry* **90,** 3156–3164 (2018).
18. Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies* **1,** a000588 (2015).
19. Barnes, S. Overview of experimental methods and study design in metabolomics, and statistical and pathway considerations. *Computational Methods and Data Analysis for Metabolomics,* 1–10 (2020).
20. Kennedy, A. D. *et al.* Metabolomics in the clinic: A review of the shared and unique features of untargeted metabolomics for clinical research and clinical testing. *Journal of Mass Spectrometry* **53,** 1143–1154 (2018).
21. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics* **8,** 1–16 (2016).
22. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11,** 98–110 (2015).
23. Wang, F. *et al.* CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Analytical chemistry* **93,** 11692–11700 (2021).
24. Wieder, O. *et al.* A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies* **37,** 1–12 (2020).
25. Zhu, H., Liu, L. & Hassoun, S. Using Graph Neural Networks for Mass Spectrometry Prediction. *arXiv preprint arXiv:2010.04661* (2020).
26. Young, A., Wang, B. & Röst, H. MassFormer: Tandem mass spectrum prediction with graph transformers. *arXiv preprint arXiv:2111.04824* (2021).
27. Ying, C. *et al.* Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* **34,** 28877–28888 (2021).
28. Park, J., Jo, J. & Yoon, S. Mass Spectra Prediction with Structural Motif-based Graph Neural Networks. *arXiv preprint arXiv:2306.16085* (2023).
29. Overstreet, R. E., King, E., Nguyen, J. & Ciesielski, D. QC-GN2oMS2: a Graph Neural Net for High Resolution Mass Spectra Prediction. *bioRxiv,* 2023–01 (2023).
30. Murphy, M. *et al.* Efficiently predicting high resolution mass spectra with graph neural networks. *arXiv preprint arXiv:2301.11419* (2023).
31. Goldman, S., Li, J. & Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks. *arXiv preprint arXiv:2304.13136* (2023).
32. Goldman, S., Bradshaw, J., Xin, J. & Coley, C. W. Prefix-tree decoding for predicting mass spectra from molecules. *arXiv preprint arXiv:2303.06470* (2023).
33. Lee, J. *et al. Set transformer: A framework for attention-based permutation-invariant neural networks* in *International conference on machine learning* (2019), 3744–3753.
34. Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nature biotechnology* **38,** 1159–1163 (2020).
35. Oono, K. & Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947* (2019).
36. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics* **8,** 1–20 (2016).
37. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
38. Domingo-Almenara, X. *et al.* The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature communications* **10,** 5811 (2019).
39. Lee, S. *et al.* BMDMS-NP: A comprehensive ESI-MS/MS spectral library of natural compounds. *Phytochemistry* **177,** 112427 (2020).
40. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478* (2021).
41. Fey, M. & Lenssen, J. E. *Fast Graph Representation Learning with PyTorch Geometric* May 2019. https://github.com/pyg-team/pytorch_geometric.

42. Morris, C. *et al. Weisfeiler and leman go neural: Higher-order graph neural networks* in *Proceedings of the AAAI conference on artificial intelligence* **33** (2019), 4602–4609.

43. Schlichtkrull, M. *et al. Modeling relational data with graph convolutional networks* in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15* (2018), 593–607.

44. Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

45. Shi, Y. *et al.* Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509* (2020).

46. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

47. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods* **13,** 741–748 (2016).

48. Fenyö, D. & Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry* **75,** 768–774 (2003).

49. Nowatzky, Y., Benner, P., Reinert, K. & Muth, T. Mistle: bringing spectral library predictions to metaproteomics with an efficient search index. *bioRxiv,* 2022–09 (2022).

50. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667 (2007).

51. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **5,** 859–866 (1994).