OXFORD

Briefings in Bioinformatics, 2025, **26(4)**, bbaf333
https://doi.org/10.1093/bib/bbaf333
Case Study

SingleFrag: a deep learning tool for MS/MS fragment and spectral prediction and metabolite annotation

Maribel Pérez-Ribera[1], Muhammad Faizan-Khan[2], Roger Giné[2], Josep M. Badia[2], Alexandra Junza[2,3], Oscar Yanes[2,4], Marta Sales-Pardo[1,*], Roger Guimerà[1,5,*]

[1]Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia
[2]Department of Electronic Engineering, IISPV, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia
[3]Servei de Recursos Científics i Tècnics, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia
[4]CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Instituto de Salud Carlos III, 28029 Madrid, Spain
[5]ICREA, 08010 Barcelona, Catalonia
[*]Corresponding authors. Marta Sales-Pardo, E-mail: marta.sales@urv.cat; Roger Guimerà, E-mail: roger.guimera@urv.cat

Abstract

Metabolite and small molecule identification via tandem mass spectrometry (MS/MS) involves matching experimental spectra with prerecorded spectra of known compounds. This process is hindered by the current lack of comprehensive reference spectral libraries. To address this gap, we need accurate *in silico* fragmentation tools for predicting MS/MS spectra of compounds for which empirical spectra do not exist. Here, we present SingleFrag, a novel deep learning tool that predicts individual fragments separately, rather than attempting to predict the entire fragmentation spectrum at once. Our results demonstrate that SingleFrag surpasses state-of-the-art *in silico* fragmentation tools, providing a powerful method for annotating unknown MS/MS spectra of known compounds. As a proof of concept, we successfully annotate three previously unidentified compounds frequently found in human samples.

Keywords: Metabolite; Machine learning; Graph neural networks; MS/MS; In silico fragmentation

Introduction

Mass spectrometry (MS) is a powerful analytical tool for studying small molecules and metabolites in biological systems. The interpretation of tandem MS (MS/MS) spectra is crucial for metabolite annotation and identification, driving advancements in metabolomics across diverse fields such as precision medicine [1], biomarker discovery [2], nutritional sciences [3], microbiome research [4], toxicology and environmental testing [5].

However, practical application of MS/MS for metabolite identification presents significant challenges. Traditional approaches rely on fragmentation libraries [6] and spectral matching, where pre-recorded MS/MS spectra of known compounds from pure standards are compared to experimental spectra to identify matching fragments [7]. The success of this method is heavily dependent on the availability of comprehensive and accurate MS/MS spectral libraries. Unfortunately, current spectral libraries suffer from limitations in size, quality, and diversity, and they are difficult to maintain and update [8, 9]. As a result, a substantial portion of MS/MS spectra generated in metabolomic experiments remains unidentified through spectral library searching methods.

To address the challenge of metabolite identification, computational tools generally adopt one of two strategies [10]: (i) computationally processing experimental MS/MS spectra and searching for putative annotations among molecules in databases of known compounds [11, 12]; or (ii) using the chemical structures of metabolites to predict their MS/MS spectra [13–18], which are then compared to experimental spectra for identification. Both approaches involve machine learning algorithms that convert experimental MS/MS spectra into feature vectors and encode chemical structures as fingerprints, embeddings, or graph structures for learning purposes.

In this study, we focus on the second approach, which aims to overcome the limitations of empirical MS/MS libraries by computationally predicting MS/MS spectra, ultimately generating reliable *in silico* MS/MS libraries. Existing tools that follow this approach employ a combination of techniques: rule-based methods that capture known fragmentation patterns based on chemical principles [13, 16], probabilistic models that assign probabilities to potential fragmentation events through statistical analysis of experimental MS/MS spectra [14, 15], and machine learning algorithms, ranging from traditional methods to deep learning [17–19].

Here, we introduce SingleFrag, a novel deep learning tool for predicting MS/MS spectra (Fig. 1). SingleFrag stands out from existing machine learning approaches in two significant ways. First, instead of predicting the entire fragmentation spectrum of a molecule with a single model, we train a separate model for each fragment ion. This approach aims to reflect that peaks (corresponding to fragment ions) in an MS/MS spectrum are not necessarily correlated. For example, one peak might correspond to a particular moiety in the molecule, while another peak corresponds to a different moiety with no relation to the first. Using the same features to predict both peaks simultaneously can reduce

Received: November 11, 2024. **Revised:** May 07, 2025. **Accepted:** May 31, 2025

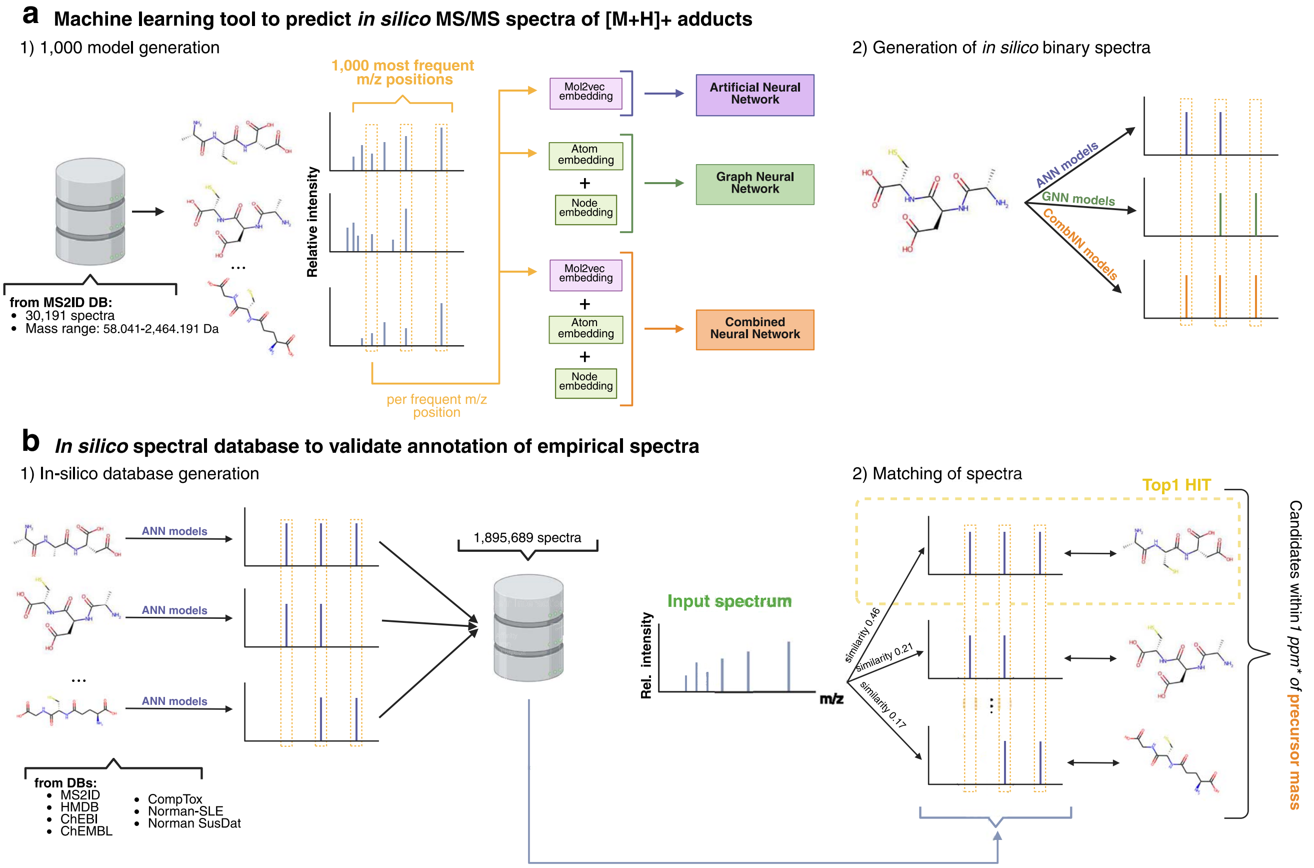
a Machine learning tool to predict in silico MS/MS spectra of [M+H]+ adducts
1) 1,000 model generation
from MS2ID DB:
• 30,191 spectra
• Mass range: 58.041-2,464.191 Da
1,000 most frequent m/z positions
Relative intensity
per frequent m/z position
Mol2vec embedding
Atom embedding
+
Node embedding
Artificial Neural Network
Graph Neural Network
Combined Neural Network
2) Generation of in silico binary spectra
ANN models
GNN models
CombNN models
b In silico spectral database to validate annotation of empirical spectra
1) In-silico database generation
ANN models
from DBs:
• MS2ID
• HMDB
• ChEBI
• ChEMBL
• CompTox
• Norman-SLE
• Norman SusDat
1,895,689 spectra
2) Matching of spectra
Input spectrum
Rel. intensity
m/z
similarity 0.46
similarity 0.21
similarity 0.17
Top1 HIT
Candidates within 1 ppm* of precursor mass

Figure 1. SingleFrag for spectral prediction and annotation. (a), We consider an empirical database of MS/MS spectra, and train three model types to predict the presence or absence of each of individual fragment ions across spectra. In particular, we build 1,000 models (of each of the three model types), corresponding to the 1000 most frequent fragment ions. Once all models are trained, we predict the whole spectrum of a given molecule by predicting each of the 1,000 peaks independently. (b), To annotate unknown empirical spectra, we build a database of *in silico* predicted spectra for over 1.8 million compounds. Then, for an unknown empirical spectrum that we wish to annotate, we select all candidates from the database with masses compatible with the unknown spectrum, and rank the candidates according to the similarity between their predicted spectra and the target empirical spectrum. Figure created in BioRender [25].

prediction accuracy. Therefore, by training individual models for each fragment, we aim to better capture the molecular and structural features associated with each specific peak. Of course, one could argue that some fragments in any given spectrum are surely related and, therefore, that using a different model for each peak may be detrimental, to the extent that these correlations are not exploited. Given these contrasting arguments, the question of whether predicting individual fragments is beneficial or detrimental can only be solved empirically, which is what we do here.

Second, SingleFrag also differs from existing machine learning approaches in that, rather than predicting the intensity of peaks in the spectrum, it focuses on predicting the presence or absence of each peak. We argue that the presence of a given fragment is more relevant for molecular structural annotation than its intensity for several reasons. First, intensity values can be highly variable due to multiple uncontrolled factors, such as ionization efficiency, instrument-specific settings, collision energy, and fragmentation technique. Modeling these values introduces additional noise and complexity, potentially leading to overfitting and reduced model generalizability across instruments and datasets. Second, our goal with SingleFrag is to capture the structural relationships between molecular substructures and their corresponding fragment ions. Presence/absence modeling aligns better with this objective, allowing the model to focus on learning the underlying chemistry of fragmentation rather than being confounded by variable intensity profiles. Third, in practice, spectral similarity metrics such as cosine similarity heavily weight intensity differences—small deviations in predicted intensities can lead to lower similarity scores, even if the predicted fragments are chemically correct. By binarizing spectra and focusing on peak presence, SingleFrag avoids this pitfall and enables more robust and interpretable matches. And finally, current public MS/MS libraries contain limited data on spectra acquired at multiple, standardized collision energies (with the exception of the NIST MS/MS library). This scarcity limits the ability of machine learning models to learn consistent intensity patterns across compounds. All in all, we argue that this approach allows for more robust and reliable predictions across different experimental conditions.

We train three SingleFrag models to predict the presence of individual fragments (Fig. 1a). The first model embeds the molecule using mol2vec [20] and uses this embedding as input to a multilayer, feedforward neural network. The second model employs a graph neural network (GNN) [21–23] that takes as input the unprocessed molecular graph. The third model integrates both the mol2vec embedding and the GNN. We intentionally keep the architecture these models simple, each model comprising only on the order of 10^3 parameters, compared to the millions of parameters of state of the art deep learning algorithms. This allows us to: (i) train models for many different peaks; (ii) assess the validity of our assumption with regards to modeling single peaks versus whole spectra (had we chosen to implement very sophisticated deep learning models, the origin of any potential improved performance of SingleFrag would remain unclear). Training three different SingleFrag models allows us to evaluate the variability of models for single peak prediction and, again, assess the potential benefits of single fragment approaches *vis a vis* whole spectrum approaches.

We find that all three SingleFrag models, particularly the first one, outperform state-of-the-art methods at predicting MS/MS fragments, including rule-based Competitive Fragmentation Modeling-ID (CFM-ID [16]) and deep learning models using transformers (MassFormer [18]) and incorporating the 3D structure of the molecule (3DMolMS [19]). Encouraged by this performance, we generate an *in silico* MS/MS library for nearly two million molecules and test its ability to identify unknown metabolites from their spectra (Fig. 1b). For each target spectrum, we rank candidates from the library based on the similarity between the target experimental spectrum and SingleFrag's *in silico* predictions for the candidates. We find that the true target molecule is ranked first in 38% of cases and within the top five candidates in 72% of cases, demonstrating high accuracy and reliability. Finally, we apply this annotation method to recurrent unidentified spectra from the ARUS database [24] of the NIST MS Data Center, successfully confirming the annotation of three metabolites by manually analyzing their molecular structures and fragmentation patterns, thereby validating the effectiveness of our approach.

Results

SingleFrag models for *in silico* prediction of individual MS/MS fragments

We develop and validate our models using a dataset containing MS/MS fragmentation spectra for 30,191 compounds sourced from Human Metabolome Database (HMDB) [26], Agilent METLIN, MassBankEU and MoNA [27, 28], NIST, and Riken [29] (Methods section). Collectively, these data bases capture the current state of accessible metabolomics data. We applied no compound class-based filtering, so that the dataset's composition reflects the underlying distribution of available MS/MS data in public reference libraries. We randomly allocate spectra from 24,473 of these compounds to the training set, 3,059 to the validation set, and 2,659 to the test set. All results reported here correspond to the test set, which is not used in any way for training or hyperparameter tuning.

In all cases, we discretize the spectra into bins of size m/z = 0.01 Da by ceiling the mass of each fragment to two decimal places. This choice is not based on *ex post* analysis of algorithm performance, but rather on a prior analysis of typical instrument precision. We find (Supplementary Fig. S1) that the measurement error of m/z is typically of the order of a few thousandths Da so making bins of order 0.001 Da would result in the same peak spreading over several bins, thus severely compromising the ability of the algorithm to learn, because each model would have access to fewer training examples. Conversely, bins of order 0.1 Da would almost invariably include several peaks corresponding to distinct fragments, again compromising the ability of algorithms to learn, in this case because each model would have to learn different mechanisms leading to the same peak.

We construct a separate model for each mass bin. Given our choice of bin size, this approach typically results in models for individual fragments, hence the name SingleFrag (however, some bins contain distinct fragments; see Methods and Supplementary Fig. S1). Building a model for every bin is computationally very demanding, so we focus on the 1,000 bins where peaks occur most frequently, thereby concentrating resources on the most informative and data-rich parts of the MS/MS spectra. Although in principle our approach is not biased toward larger or smaller fragments, smaller fragments tend to be more common. Therefore, the bins we consider end up covering masses m/z in the range [29.04, 269.09], and account for 60% of all peaks in the spectra in our dataset (Fig. 2a). By restricting spectra to these 1,000 bins, 94% of the compounds in the dataset still have 3 or more peaks, 90% have 5 or more, and 80% have 10 or more (Fig. 2b).

Each of the 1,000 models is a binary classifier that takes as input the molecule and predicts whether a peak exists in the corresponding mass bin. As discussed earlier, we focus on binary predictions because the presence of a given fragment is more relevant and robust for molecular structure annotation than its intensity, which can vary based on factors such as collision energy or the mass spectrometer used to obtain the spectrum. To account for all possible fragments associated with a molecule, we consider all available empirical MS/MS spectra for a given molecule and build a single merged discretized spectrum with binary values, so that m/z bins in which a peak is present in at least one spectrum are equal to one, and other bins are equal to zero (Methods section).

We investigate three different SingleFrag models as our peak binary classifiers (see Methods section for details). First, we use a multilayer feedforward artificial neural network (ANN). In this model, molecules are embedded into a 300-dimensional space using the previously trained mol2vec model [20], and the resulting embeddings are fed into the neural network. Second, we employ a GNN that directly uses the molecular graph as input [21–23]. Finally, we combine these approaches into a model that uses both the GNN and the mol2vec embeddings. Each SingleFrag model is simple, comprising only a few thousand parameters (compared to millions in whole-spectrum models like MassFormer or 3DMolMS). This makes training a model for a given fragment very fast, and allows us to train models for the 1,000 fragments discussed above (for each SingleFrag model type: ANN, GNN, and combined), with a cost that is comparable to that of training existing large models for whole-spectrum prediction.

Each SingleFrag model returns a score between 0 and 1 for each input molecule and m/z bin. To evaluate the model's performance, we convert these scores into binary predictions for the presence (1) or absence (0) of a peak. To this end, we calculate a specific threshold for each m/z bin, so that scores above the threshold are converted into a 1 and scores below the threshold are converted into to a 0. We select this threshold so that predictions are well calibrated for the validation set, that is, that the fraction of molecules predicted to have a specific peak matches the fraction of molecules with that peak in the training set (see Methods section for details).

Figure 2c–e illustrates the performance of each model type (ANN, GNN, and combined) across different mass bins (see also Supplementary Fig. S2 for details about each individual ANN model). We evaluate performance using accuracy and the F1 score (see Methods section for details). Due to the unbalanced nature of the target values—where even the most frequent peaks are relatively rare—accuracy often approaches 1 and does not provide a clear picture of model performance. In contrast, the F1 score, which is close to both precision and recall because our models are well-calibrated, is more informative. Our results show that the ANN and combined models, with mean F1 scores of 0.411 and

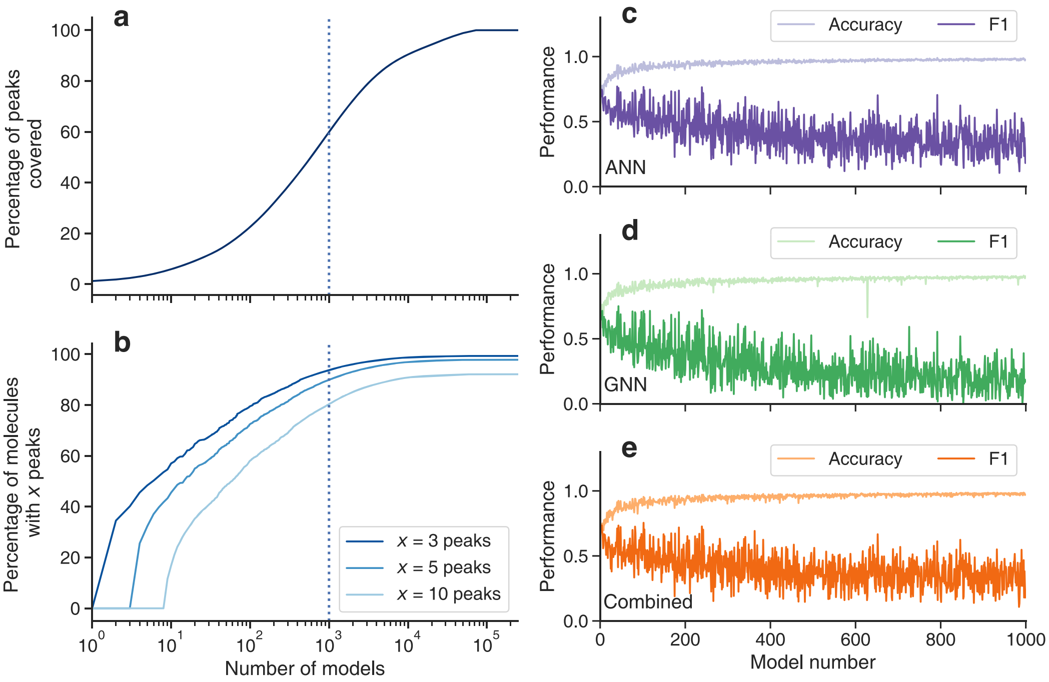

Figure 2. Prediction of individual fragments. We discretize the m/z axis of each spectrum into bins of width 0.01. (a) The percentage of peaks in the training set covered by modeling only the x bins with the highest frequency of peaks across molecules (x-axis). The dashed line indicates the coverage achieved by 1000 models. (b) The percentage of molecules in the training set whose spectra contain 3, 5, and 10 peaks when only the x most frequent bins are considered. The dashed line indicates the coverage achieved by 1000 models. Based on (a) and (b), we select the 1,000 bins that cover the highest fraction of peaks and molecules. These bins (a) cover 60% of the peaks in the training set and (b) ensure that 94% of the training spectra have at least 3 peaks, 90% have at least 5 peaks, and 80% have at least 10 peaks. For each of these bins, we train three different SingleFrag models (ANN, GNN, and combined; see text for details). (c to e) Accuracy and F_1 score for the each bin and model type: (c) ANN, (d) GNN, and (e) combined. The models on the x-axis are sorted from highest to lowest frequency of peaks in the training set, with the first model corresponding to the bin that most frequently contains a peak (at m/z 51.03) in the training set.

0.399 respectively, outperform the GNN model, which has a mean F1 score of 0.290. The good performance of the ANN model rests on the fact that molecules with similar mol2vec embedding also tend to have more similar spectra (Supplementary Fig. S3).

Our results also show that, on average, prediction accuracy decreases for increasingly rare peaks, which provides a rationale for limiting the number of peaks considered in our approach. Other peaks (including those corresponding to larger fragments, which may be structurally informative for specific complex molecules) are typically difficult to predict because no machine-learning algorithm can learn from very few instances.

SingleFrag models yield accurate *in silico* spectra

Next, we investigate whether SingleFrag models, trained to predict individual peaks, can accurately predict whole spectra. For this, we apply all 1,000 bin-specific models to each molecule in the test set to obtain the corresponding predicted spectrum (Fig. 3). We perform this analysis for each type of SingleFrag model (ANN, GNN, and combined; Fig. 3d–f). To evaluate the performance of the SingleFrag models, we benchmark them against three state-of-the-art algorithms: (i) CFM-ID [16] (Fig. 3b), the leading rule-based prediction algorithm, and widely used for *in silico* fragmentation; (ii) 3DMolMS [19] (Fig. 3a), which takes into account the 3D structure of molecules in its predictions; and (iii) MassFormer [18] (Fig. 3c), which uses a graph transformer architecture to model relationships between atoms in the molecule.

We compare the predicted spectra in terms of precision, recall, accuracy, and cosine similarity between the predicted and real spectra in the test set (Fig. 3h–j). These performance metrics vary considerably across compounds (although we do not observe any clear dependency of performance on molecular properties such as molecular size; see Supplementary Fig. S4). Because of this, and given that our validation naturally provides matched samples (predictions from different algorithms for each compound), we report the number of molecules for which each tool has the best metric. This comparison is restricted to molecules that could be predicted by all methods, specifically those for which CFM-ID, 3DMolMS and MassFormer return valid outputs, accounting for 2,563 compounds in the test set. As before, we disregard the intensity of fragment ions and focus solely on whether peaks are predicted to exist or not.

Even though predictions are obtained at the provided m/z resolution for each method (0.01 Da for SingleFrag, 0.2 Da for 3DMolMS, and 1 Da for MassFormer), and to ensure consistency, we calculate performance metrics using m/z bins of size 1, thus matching the lowest precision provided by all methods (MassFormer). Additionally, we compare whole spectra, and not only the 1,000 fragments that SingleFrag has been trained to predict—all other fragments are predicted not to exist by SingleFrag, and counted as errors if they actually exist in the empirical spectrum. Finally, note also that, although no molecule in the test set has been used in any way for training SingleFrag, some test molecules could be in the training set of benchmark algorithms.

As shown in Fig. 3h–j, we find that 3DMolMS performs better than CFM-ID and MassFormer. 3DMolMS and CFM-ID perform similarly in terms of precision, but 3DMolMS tends to have higher recall, resulting in higher cosine similarities between the predicted and real spectra. MassFormer performs similarly to, but always slightly worse than 3DMolMS. However, all three SingleFrag models achieve better precision than CFM-ID, 3DMolMS and MassFormer. Since their recall is comparable to 3DMolMS (better for the combined model, similar for the ANN model,

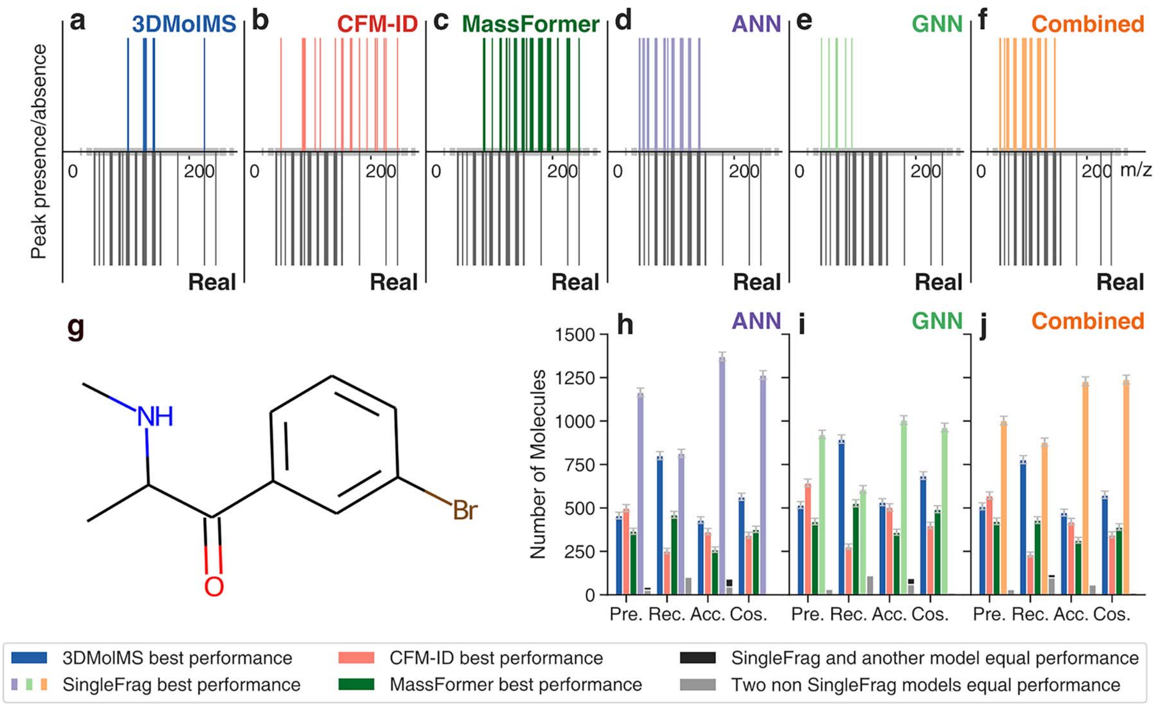

Figure 3. Prediction of whole spectra. (a–f) We predict the MS/MS spectrum of the molecule depicted in (g using: (a) 3DMolMS [19]; (b) CFM-ID [16]; (c) MassFormer [18]; (d) SingleFrag ANN model; (e) SingleFrag GNN model; and (f) SingleFrag combined model. As throughout the paper, we disregard the intensity of peaks and just represent their presence or absence. Note that, by construction, SingleFrag models cannot predict all peaks, but only those at the 1,000 most frequent bins (grey ticks in the m/z axis; see text for details). However, when comparing to the real spectra, all m/z positions are considered. (h–j) We benchmark the performance of SingleFrag models against 3DMolMS, CFM-ID and MassFormer. For each molecule in the test set, we calculate the precision, recall, accuracy, and cosine similarity of the predicted spectrum compared to the real spectrum. Since these performance metrics vary considerably across molecules and our validation naturally gives matched samples (that is, for each molecule we obtain the spectrum predicted by each of the algorithms), we plot, for each metric, the number of molecules for which a given model is the best performer. Error bars represent the standard deviation obtained by boostrapping.

and worse for the GNN model), SingleFrag models produce *in silico* spectra with higher accuracy and cosine similarity to the true spectrum overall. This is particularly noteworthy considering the factors outlined above and playing against SingleFrag, namely: (i) ignoring its higher resolution to match the lower resolution of the benchmarks; (ii) not limiting the comparison to the 1,000 trained m/z bins; and (iii) probably including in the test set molecules on which the other algorithms (but not SingleFrag) were trained. The fact that, despite all this, the predictions of all SingleFrag models are, overall, closer to the true spectra than those of the benchmarks strongly supports our hypothesis that modeling individual peaks separately results in more accurate predictions than using a single model for the entire spectrum.

A database of *in silico* spectra of known molecules enables the annotation of unknown empirical spectra

Given the success of SingleFrag models, particularly the ANN model, in predicting whole spectra, we investigate whether these *in silico* predicted spectra are accurate enough to annotate unknown MS/MS empirical spectra of known compounds. These are compounds that have been described in chemical databases but whose fragmentation spectra are not available to the community. To facilitate this annotation process, we have created a database containing nearly 1.9 million *in silico* MS/MS spectra predicted by the ANN SingleFrag model. To generate these spectra, we begin with the SMILES representations of nearly 1.9 million compounds (Methods section). Using these SMILES, we first obtain their 300-dimensional mol2vec embeddings, and then perform a forward pass through the SingleFrag ANN model to predict their spectra.

With the resulting database, we can annotate unknown spectra through the following steps. Given an unannotated empirical spectrum, we first estimate either: (i) the molecular mass of the corresponding compound based on the mass of the precursor ion, or (ii) its exact molecular formula using a tool such as BUDDY [30]. Next, we filter the database of 1.9 million compounds/spectra to find candidate compounds that match the mass or molecular formula, respectively, of the unknown target compound. Finally, we rank these candidates by calculating the cosine similarity between the unannotated empirical spectrum and the *in silico* spectrum predicted for each candidate.

To validate the performance of SingleFrag at the task of annotating spectra, we use the same test set of molecules as in previous sections. For each MS/MS spectrum, we generate a list of candidate molecules with compatible molecular formulas (Fig. 4a and d) or compatible molecular masses, assuming a spectrometer precision of either 1ppm (Fig. 4b and e) or 10ppm (Fig. 4c and f). We evaluate the annotation performance by tracking the position of the true molecule in the candidate ranking for each test molecule. In Fig. 4a-c, we show the frequency with which the true molecule is ranked first (Top 1, indicating perfect annotation), among the top five candidates (Top 5), and among the top ten candidates (Top 10). We measure this performance under four experimental conditions: using low (0–14 eV), medium (15–39 eV), or high (≥40 eV) collision energy spectra as the unknown empirical query (800, 1057, and 542 spectra, respectively, from the

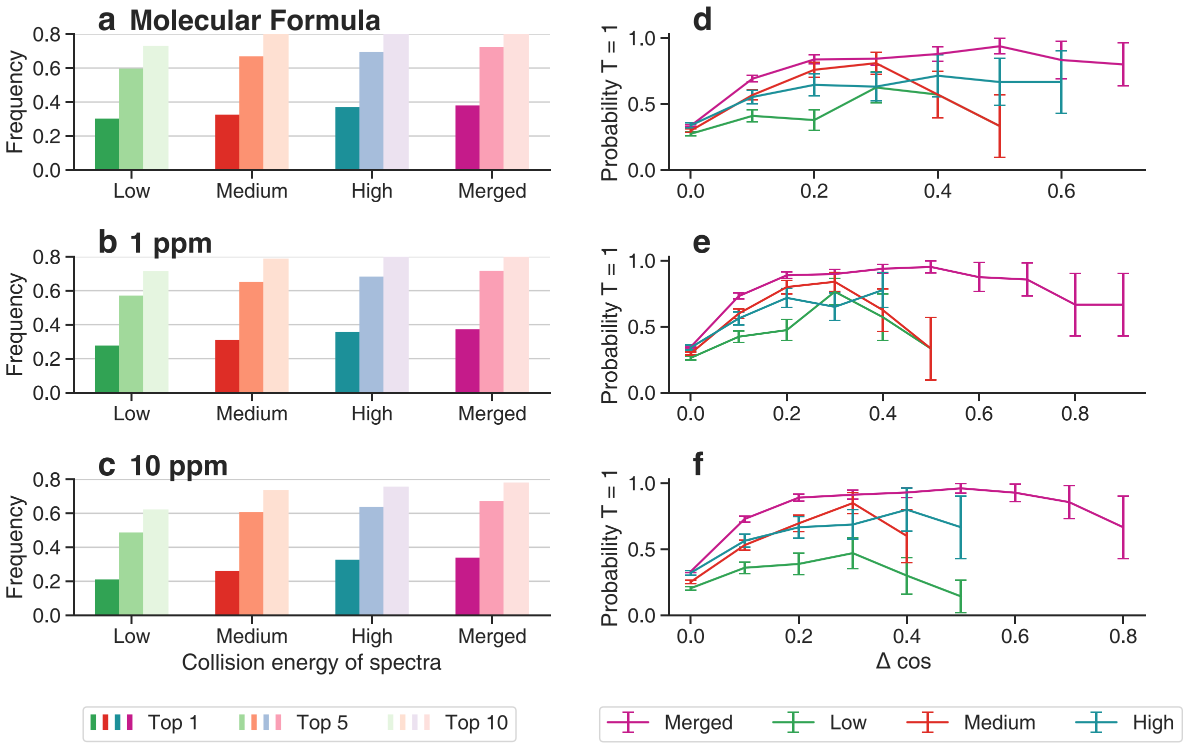

Figure 4. Annotation of unknown empirical MS/MS spectra. We evaluate the ability of SingleFrag to annotate empirical MS/MS spectra of known compounds, meaning compounds that are described in databases but whose reference fragmentation spectra are not available. For each molecule in the test set, we identify putative annotations by selecting compatible candidates from a reference database containing nearly 1.9 million known molecules. Compatibility is established based on: (a, and d) exact molecular formula (simulating a scenario where the exact formula of the test molecule is determined with a tool such as BUDDY [30]); (b and e) exact mass, with a window of ±1 ppm; or (c and f) exact mass, with a window of ±10 ppm. We then rank candidate annotations by computing the cosine similarity between the empirical spectrum of the test molecule and the predicted SingleFrag ANN spectrum of each compatible candidate. For each test molecule and experimental condition (exact formula, 1 ppm, and 10 ppm), we consider separately empirical spectra with different collision energies: (i) low collision energy (<15 eV); (ii) medium collision energy (≥15 eV and <40 eV); (iii) high collision energy (≥40 eV). Additionally, we consider merged spectra processed as in the training of SingleFrag, by merging spectra for all collision energies available for each test molecule. (a–c) We plot the frequency with which the true molecule (Tanimoto coefficient [31] $T = 1$) is ranked as the top candidate (perfect annotation), and among the top 5 and top 10 candidates (plausible annotations). (d–f) For each experimental condition, we plot the probability of the top candidate being the true molecule (Tanimoto coefficient [31] $T = 1$) as a function of the score gap Δ cos between the top and the second ranked candidates. Generally, larger gaps provide more confidence in the top annotation.

NIST20 and Agilent METLIN Metabolomics databases; Methods section), and using merged query spectra where peaks at all available collision energies are combined into a single spectrum (see Methods section for details).

We obtain the best results when we use as much information as possible, that is, when the exact molecular formula of the target molecule is assumed to be known from a tool such as BUDDY [30], and multiple spectra (at different collision energies) for the molecule are merged (Fig. 4a). Under these conditions, the true molecule is ranked first in 37.6% of cases, among the top five candidates in 71.9% of cases, and among the top ten candidates in 82.7% of cases. Performance decreases only slightly when individual high-energy spectra are used (36.6% ranked first, 69.0% top five, 81.2% top ten). Performance decreases more noticeably when medium-energy spectra (32.2% ranked first, 66.5% top five, 79.9% top ten) and low-energy spectra are used (29.9% ranked first, 59.3% top five, 72.6% top ten). This happens because SingleFrag focuses on the 1,000 bins where peaks occur most frequently, covering masses in the low mass range between m/z 29.04 and 269.09. These bins are more likely to be present in high-energy spectra, which typically produce extensive fragmentation, especially in the low mass range. By contrast, low-energy spectra tend to be less fragmented, dominated by a few large, rare fragments or the molecular ion itself, for which SingleFrag is likely to not make predictions.

Significantly, in all cases, the reliability of annotations can be considerably increased by examining the relative position of the candidates in the ranked list (Fig. 4d-f). As mentioned, we score candidates based on the cosine similarity between the target empirical spectrum and the *in silico* predicted spectrum for each candidate. However, the difference between the scores of the first and second candidates is particularly informative about the reliability of the top candidate. When the two scores are very close, the top candidate is the true metabolite in 20%–30% of cases (Fig. 4d–f), consistent with the overall figures in Fig. 4a–c. However, when there is a large gap between the cosine similarities of the first and second candidates, the top candidate is much more likely to be the true molecule. Specifically, when the difference in cosine similarity scores (Δ cos) between the top and second ranked candidates exceeds 0.2, the top candidate is the correct molecule in the majority of cases. (This is specially true for merged spectra and at high collision energies; at low energies, the small number of instances with gaps larger than 0.4 sometimes obscures this signal.)

In less clear situations, when candidate lists have smaller gaps between top candidates, SingleFrag is still useful in annotating, because smaller gaps are often related to top candidates being structurally similar to each other. For example, when the gap is $\Delta \cos < 0.1$, the top two candidates have an average Tanimoto similarity ranging from $T = 0.39$ in the worse case, when using low-energy spectra and a 10ppm window for candidates, to $T = 0.52$ when using merged spectra and the exact molecular formula for candidates. These values are significantly larger than the null expectation $T = 0.17 \pm 0.10$ for the Tanimoto coefficient of randomly selected pairs of molecules in our dataset. Thus, even when the top candidate does not coincide with the ground-truth molecule, it provides structural cues about its structure.

Annotation of recurrent unidentified spectra from the ARUS database

Building on the validation results from the previous section, we demonstrate that our approach effectively annotates unknown spectra. We focus on spectra from the Annotated Recurrent Unidentified Spectra (ARUS) database [24], maintained by the NIST MS Data Center. This database includes spectra that frequently appear in real samples but remain unannotated.

We used a dataset of ARUS spectra with putative molecular formulas assigned using BUDDY [30]. The dataset includes spectra from two sources: plasma (25,801 spectra) and urine (68,478 spectra). For each unknown molecule, we discretized and binarized its associated spectrum as previously described. We then generated a list of candidate annotations for each spectrum using a filtering window of 1 ppm. To be as exhaustive as possible in the annotation of these unidentified spectra, we enlarged the *in silico* database with all compounds available for download from PubChem, resulting in an extended database of over 96 million predicted spectra. We ranked the candidate annotations using the similarity between the target empirical spectrum and the SingleFrag *in silico* prediction for each candidate's spectrum and kept the ten candidates whose *in silico* spectra were more similar to the query spectrum.

From hundreds of potential annotations, and given that validating annotations manually is very costly, we selected three particularly promising ones. This selection was based on two criteria: (i) the molecular formula matched BUDDY's prediction, with a low estimated false discovery rate in BUDDY; and (ii) there was a significant gap between the first and second candidates in SingleFrag's ranking. We confirmed the annotation for these three compounds—glucuronyl-2-hydroxyhippurate, 8-hydroxyquinoline glucuronide, and a truxilline isomer—through the manual elucidation of their fragment ions from their molecular structures by an expert chemist (Fig. 5). These three compounds are of particular interest due to their putative origin. Isomeric truxillines, a group of minor alkaloids consistently found in illicit cocaine samples, suggests possible exposure through drug use or environmental contamination related to cocaine production. 8-hydroxyquinoline glucuronide, a metabolite of 8-hydroxyquinoline, could indicate pharmaceutical use or exposure to quinoline compounds. Glucuronyl-2-hydroxyhippurate, a conjugate of hippuric acid, may reflect dietary intake or endogenous metabolic processes involving aromatic acids.

Discussion

SingleFrag is a novel *in silico* fragmentation tool for metabolites and small molecules. Unlike existing machine learning tools, SingleFrag focuses on predicting the presence or absence of individual fragments rather than attempting to predict whole spectra, including all peaks and their intensities. Although it can be argued that some peaks are related, and that ignoring these relationships may lead to less accurate predictions, our approach is based on the rationale that the molecular features predictive of the existence of a specific fragment may be distinct from those predictive of most other fragments. Our results validate this rationale.

We developed three different SingleFrag models: ANN, GNN, and combined. While sharing the idea of modeling the existence of single fragments, these models operate in fundamentally different ways. The ANN model uses a mol2vec embedding, while the GNN model works directly with the molecular graph, leading to distinct neural network architectures and underlying mathematical models. Despite these differences, all SingleFrag models outperform the state-of-the-art *in silico* prediction tools, CFM-ID, 3DMolMS and MassFormer. This is particularly noteworthy for MassFormer, which employs a more sophisticated network architecture, and for 3DMolMS, which also uses a much larger architecture and, additionally, takes into account detailed features encoding the 3D location of each atom in molecules.

A consequence of SingleFrag's approach is the need to train a different model for each fragment we aim to predict. Here, we have shown that 1,000 such models are enough to produce *in silico* spectra that are more accurate than the state of the art in whole-spectrum approaches, but two reasonable objections can be made. First, one may argue that some rare fragments may be very informative for certain molecules, and will be missed by our approach. Second, one may add that considering more models to alleviate the first objection is not scalable and that, in any case, training 1,000 models may become prohibitive if the size of spectral databases grows. With respect to the first objection, we argue that rare fragments may by structurally informative but, in general, will not be very useful to machine learning approaches, which typically cannot learn from only a few examples. This argument is supported by the fact that, even within the 1,000 fragments considered here, the most common ones are better predicted than those that are less frequent (Fig. 2). Additionally, we note that adding more fragments does not seem to add to SingleFrag's performance—we explored the addition of another 1,000 models, but observed no significant improvement in performance. With regards to the second objection, we note that each SingleFrag model consists of only a few thousand parameter, as opposed to the millions of parameters used in whole-spectrum approaches. Therefore, training 1,000 SingleFrag models is not more costly than existing approaches.

All things considered, we hope that our approach of modeling peaks individually will inspire new developments where more complex features and models, such as those used in 3DMolMS and MassFormer, can be integrated into SingleFrag-like frameworks to achieve even higher levels of spectral predictive accuracy. Additionally, we recognize that there is significant potential for enhancing the neural network architectures we have employed. Indeed, we deliberately kept our models simple to demonstrate that the improved performance is primarily due to the approach of modeling individual peaks instead of whole spectra, rather than more sophisticated deep learning techniques. This is especially relevant for the GNN model. Its slightly lower performance compared to the ANN model should not be seen as a fundamental limitation of GNNs in the domain of spectral prediction, but rather as an opportunity to refine graph features and optimize graph layer architectures for even better results. Our current approach considers only very basic atom and bond featurization—more sophisticated featurizations of atoms and bonds, or including graph attributes and molecular fingerprints, are likely to improve GNN performance.

We have also shown that, even with relatively simple neural network models, SingleFrag can already be used to annotate real spectra, both in a controlled setting with test molecules,

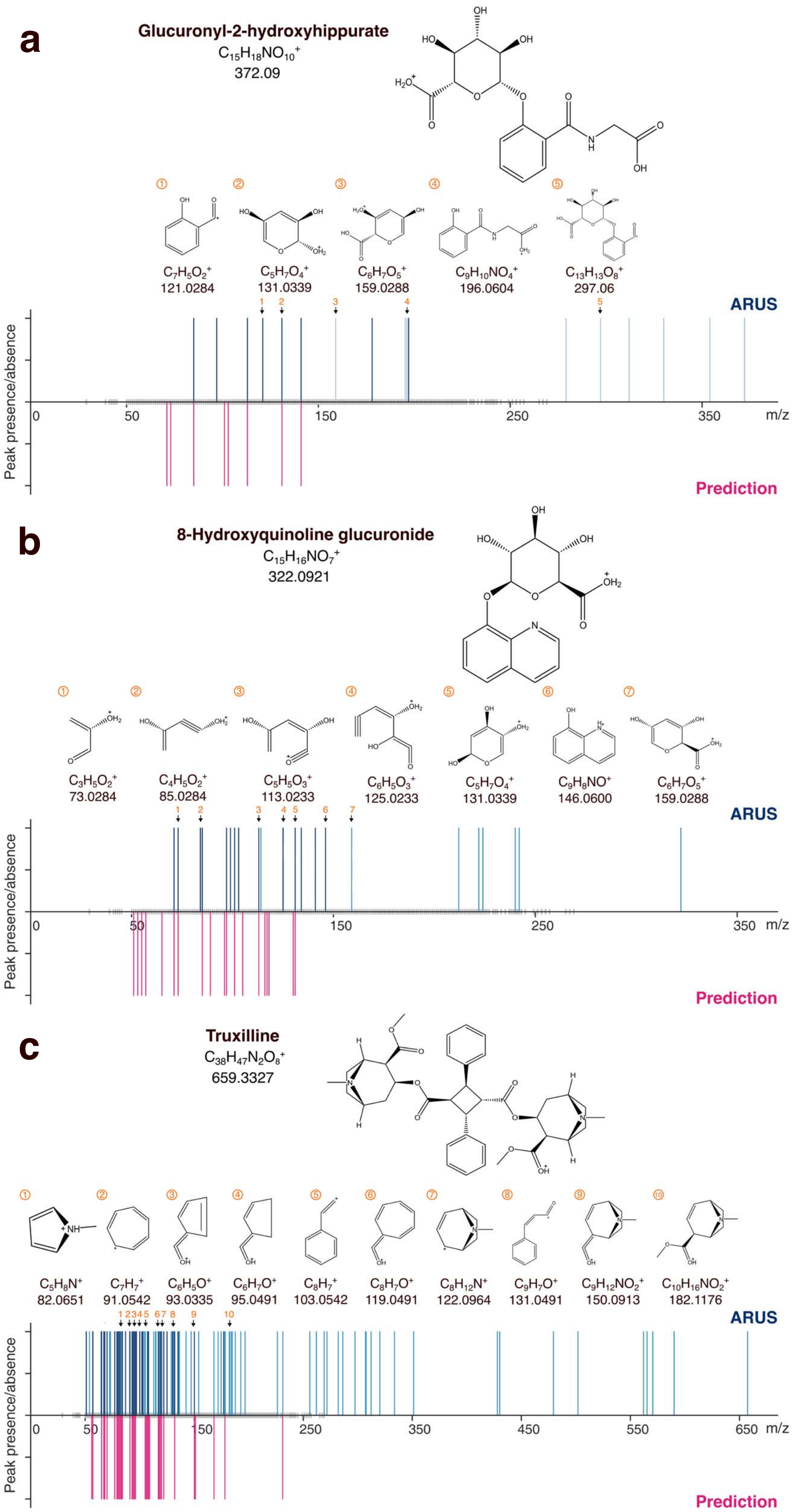

Figure 5. Annotation of three empirical MS/MS spectra from the ARUS database [24]. For three different unannotated spectra in ARUS, we show the empirical MS/MS spectrum (obtained by merging all available spectra for the compound and binarized as in the rest of the paper) and the predicted spectrum for the top candidate annotation identified by SingleFrag. In each case, we show the chemical structure of the top candidate, as well as the match between observed peaks of the empirical spectrum and plausible fragments identified manually via structural analysis of the molecule.

and in a real-world scenario with ARUS molecules that we have annotated for the first time. We expect that future SingleFrag-like algorithms will enable even more reliable annotation, which is the standing roadblock for the development of the full potential of metabolomics.

Moreover, our results emphasize the critical importance of acquiring experimental MS/MS spectra for each precursor ion at multiple collision energies. Among the widely used spectral libraries, only the NIST MS/MS provides a systematic and curated repository of spectra acquired at multiple collision energies. By contrast, other spectral libraries such as GNPS, MSDial, or MassBank often lack sufficient curation and typically contain spectra acquired at only one or a few collision energies per compound. This limitation can lead to incomplete or less informative spectra, which may not fully represent the diversity of fragment ions produced under different collision conditions.

By employing techniques like stepped collision energy, it is possible to capture a broader range of fragment ions, resulting in more detailed and informative spectra. Our results suggest that this approach may be essential, not only for real metabolomic experiments, but also for the expansion and curation of existing reference spectral libraries. This creates a positive feedback loop—improved predictive models generate better *in silico* spectra, which in turn can more accurately match the experimental spectra of unidentified compounds. This enhances the reliability and utility of these predictions, ultimately leading to new discoveries and a more comprehensive understanding of metabolic pathways and processes.

Methods

Machine learning tool to predict *in silico* MS/MS spectra of [M+H]+ adducts

Data We obtained 2 291 119 MS/MS spectra from the HMDB, Agilent METLIN, MassBankEU and MoNA, NIST, and Riken databases, corresponding to 478 631 unique compounds. The annotations for these spectra and compounds may include the name, molecular formula, InChIKey, SMILES, adduct, precursor ion mass, collision energy, and mass spectrometer type.

After excluding *in silico* generated spectra (mainly from HMDB), we retained 450 248 experimental MS/MS spectra for the protonated adduct in positive mode ([M+H]). Limiting our analysys to this adduct is based on both data availability and model robustness. The [M+H]+ adduct is by far the most commonly encountered ionization product across compound classes and instrumental platforms. While some databases do provide spectra for alternative adducts, these are relatively sparse, and their fragmentation patterns are often less reproducible or informative. We discarded spectra with m/z values reported with a single decimal or generated by low-resolution instruments. After filtering, we had 434 480 spectra corresponding to 54 790 compounds. We then merged all spectra corresponding to each unique compound into a single spectrum, regardless of the original database or acquisition tool. Using unique database identifiers for each molecule, we combined the m/z values from all corresponding spectra into a single binary list (see below details about how we build binned spectra). To ensure consistency, we obtained canonical SMILES from the latter via their InChIKeys and the Chemical Identifier Resolver API service of PubChem (https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/inchikey/). After removing 11 916 molecules without InChIKeys and 12 282 molecules not found by the request, we obtained a final dataset of 30 191 unique compounds.

We ensured that each molecule in the dataset had its exact mass, verifying that it matched the mass calculated from its SMILES within a 0.01 difference. Only one molecule did not match, so we eliminated it to avoid inconsistencies. We also verified that each spectrum contained the m/z of the precursor ion corresponding to the [M+H] adduct, allowing a window of ±0.05 Da. For spectra without such a precursor ion, we manually added it by summing the mass of a proton to the neutral mass of the compound. Additionally, we removed the peaks for m/z values larger than the mass of the precursor ion plus 0.05. This process resulted in a dataset with molecules having masses between 58.041 and 2464.191 Da. Finally, we randomized the data and divided it into a training set containing 24 472 spectra, a validation containing 3059 spectra, and test sets containing 2659 spectra.

Spectrum binarization and merged spectrum construction SingleFrag predicts the existence/absence of peaks in m/z bins with a resolution of 0.01 m/z. To that end we discretize the 450 248 spectra in our database by assigning peaks to the m/z bin corresponding to the ceiling of the second decimal of the m/z value corresponding to the peak. For instance, if a spectrum has a peak (fragment) at m/z 110.253, we assign it to m/z bin 110.26. Because, we binarize spectra, m/z bin 119.26 would have associated a value of 1 since there is a peak in that bin. At the end of the process, the binarized, discrete spectrum is a vector in which a value of 1 indicates the presence of a peak in the corresponding bin and a value of 0 indicates the absence of a peak in that m/z bin.

To obtain the merged spectrum of each unique compound, we consider all individual spectra for that compound. For each m/z bin, we look at whether individual spectra have a peak in that bin. If at least one spectrum has a peak, we assign a 1 to that bin, and a 0 otherwise. Therefore the merged spectrum of a compound represents all the possible fragments that have been detected for that compound.

Note that binning spectra is a necessary choice when using machine learning methods with discrete inputs. As explained in the main text, by using 0.01 m/z bins and using the 1000 most commons bins with associated fragments, we have a good coverage of the fragments observed in the majority of spectra. We note that, by binning spectra into 0.01 m/z bins, nonzero intensities associated to the same fragment may be split into two neighboring bins—for instance, fragments with m/z values 25.889, 25.891 would be classified in different bins, but could have been attributed to the same bin had we binned spectra differently. To assess how often this happens, we chose 100 random bins of size m/z 0.01 and represented the distribution of fragments associated to that bin (Supplementary Fig. S1). Our results show that approximately half of the time the fragments were distributed in the middle of the bins, and in the other half the fragments were distributed between neighboring bins, which suggests that there is no perfect binning strategy.

Machine learning models We introduce three SingleFrag models. All models aim to predict whether a molecule has a peak at a specific mass bin or not, and are thus binary classifiers:

1. **ANN**: First, we create 300-dimensional embeddings for all molecules in the training, validation and test sets using the mol2vec algorithm [20]. These embeddings were then input into a Multilayer Perceptron with three fully-connected layers, two of them with ReLU activation functions, and the final layer with a Sigmoid activation. The model was trained with a batch size of 16, a learning rate of 1×10^{-4}, and binary cross entropy as the loss function. We set a minimum of 200 and a maximum of 2000 epochs for training, and keep the model with the lowest validation loss. We also implemented

an early auto-stop if the validation loss increased while the training loss decreased, comparing the most recent epoc to the previous 100 epochs.

2. **GNN**: The GNN model uses the whole molecular graph to make predictions. We used bond embeddings (type of bond and atom indices) and atom embeddings (neighboring atom sequence, total hydrogens, formal charge, mass, and aromaticity). We run the node embeddings through three graph attention layers (GAT; these assign different importance to each connection in the graph) with ReLU activations. After the third GAT layer, we aggregated embeddings using sum, max, and average functions, and used a sigmoid as an activation function. We used training parameters equal to those of the ANN model but with a minimum of 4000 and a maximum of 10 000 epochs.
3. **Combined Neural Network (CNN)**: This model integrates the ANN and GNN models by using both the 300-dimensional mol2vec embeddings and the graph embeddings. The first three layers of the CNN matched the GNN structure, using graph attention operators and ReLU activations, followed by pooling. To integrate the GNN and ANN predictions, we reshaped the mol2vec vector dimensions to align with the pooled data. The ANN component included a multilayer perceptron with three fully-connected layers, two ReLU activations, and one sigmoid activation. After combining these outputs, we applied a linear layer with 45 neurons and another linear layer with ReLU, followed by a final linear layer with a sigmoid activation. Training parameters were consistent with the previous models, with a batch size of 16, a learning rate of 1e-4, binary cross entropy loss, and a training range of 4000 to 10 000 epochs.

Threshold scores and model calibration To produce *in silico* binary spectra, we need to convert the output scores from our machine learning models (ANN, GNN, and CNN) into a 0 or 1 prediction for the existing of each peak. A way to achieve this is to specify a threshold value for each model associated to a m/z bin so that scores above the threshold are converted into a 1 (presence of a peak) and the remaining scores are converted into a 0 (absence of a peak).

Because the training and test sets are selected at random, the expected fraction of molecules having a peak in a specific m/z bin in the train, validation and test sets is the same. Therefore we have to select a threshold that recovers the statistically correct fraction of molecules with peaks in that bin. To that end, for a specific m/z bin, we rank the N_{val} molecules in the validation set according to their score for that bin. We then compute the fraction $f_{\mathrm{m/z}}$ of molecules with a peak in that bin within the training set. We set the threshold as the score of the molecule in position $f_{\mathrm{m/z}}N_{\mathrm{val}}$ within the validation set (rounded upwards to the next integer). Note that the threshold value depends on the modeling approach (ANN, GNN, and CNN) and the m/z bin we consider. By binarizing scores in this way, we ensure that our models are statistically calibrated.

Validation To evaluate our predictions, we compare them with those of three other *in silico* fragmentation tools: CFM-ID [16], 3DMolMS [19] and MassFormer [18].

CFM-ID uses CFM and machine learning to fit model parameters. The latest version, CFM-ID 4.0, is accessible via a web server at http://cfmid4.wishartlab.com/ and as downloadable Docker images at https://hub.docker.com/r/wishartlab/cfmid [16]. Using CFM-ID 4.0 with its default values, which exclude fragmentations below a 0.001 probability threshold, we obtained predictions for 2638 out of the 2659 molecules in the test set. The tool CFM-ID 4.0 calculates spectra for low (10 eV), medium (20 eV), and high (40 eV) collision energies, representing them as lists of mass-intensity pairs, each corresponding to a peak in the spectrum. We then consolidated the predicted spectra at different collision energies into a single merged spectrum and removed the peaks whose intensity was lower than 1%.

3DMolMS uses deep learning to predict MS/MS spectra from 3D molecular conformations and other molecular features. The source code for the tool is available from https://github.com/JosieHong/3DMolMS. We used the pretrained model provided in this repository, and generated *in silico* spectra at low (10 eV), medium (20 eV), and high (40 eV) collision energies. We then consolidated the spectra predicted at different collision energies into a single merged spectrum and removed the peaks whose intensity was lower than 1%. 3DMolMS produces spectra at 0.2 m/z resolution.

MassFormer uses a graph transformer architecture to model long-distance relationships between atoms in the molecule. We used their pretrained model from https://github.com/Roestlab/massformer to make predictions. As we did before, we merged spectra predicted at different collision energies (normalized collision energies of 20, 40, 60, 80, and 100%) into a single merged spectrum. In this case, we removed the peaks in each spectrum whose intensity was lower than 1% of the highest predicted intensity.

Ultimately, 2,563 spectra were predicted using CFM-ID 4.0, 3DMolMS, MassFormer, and SingleFrag. Because MassFormer produces spectra at a 1 m/z resolution, we discretized CFM-ID 4.0, 3DMolMS, and SingleFrag spectra at the same resolution. We also consider only binary spectra and do not use information about the intensities associated to m/z bins or fragments.

To evaluate our models, we calculated various metrics comparing the real MS/MS spectra of the molecules (discretized and binarized) with their *in silico* spectra reconstructed by different methods (ANN, GNN, Combined, CFM-ID, 3DMolMS, and MassFormer). The metrics used were precision, recall, accuracy, F1 score, and cosine similarity. To define these metrics, we use the following nomenclature: true negative (TN), true positive (TP), false positive (FP), and false negative (FN). A model prediction is Positive if it is 1 and Negative if it is 0. Additionally, a prediction is True if it is correct and False if it is incorrect.

- **Precision**: Measures the proportion of true positive predictions among all positive predictions made by the model. It indicates how many of the predicted positives are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

- **Recall**: Also known as sensitivity, it measures the proportion of true positive predictions among all actual positives. It indicates how well the model can identify positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

- **F1**: Combines Precision and Recall into a single metric by taking their harmonic mean. It provides a balance between Precision and Recall, especially useful when there is an uneven class distribution.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- **Accuracy**: Measures the proportion of all correct predictions (both true positives and true negatives) among the total number of cases evaluated. It indicates the overall effectiveness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- **Cosine Similarity**: Measures the cosine of the angle between two vectors in an inner product space. It quantifies how similar the vectors are by determining if they point in approximately the same direction.

$$\cos(A, B) = \frac{A \cdot B}{\|A\| \ \|B\|} \tag{5}$$

In silico spectral database to validate annotation of empirical spectra

Data To build the *in silico* spectra database, we used the previously compiled list of SMILES of 30 191 unique compounds with associated experimental spectra. We also downloaded SMILES of metabolites from various online sources: HMDB (217 760 SMILES), Chemical Entities of Biological Interest (ChEBI) (51 112 SMILES), ChEMBL (1 773 996 SMILES), Computational Toxicology (CompTox) (8963 SMILES), NORMAN Suspect List Exchange (NORMAN-SLE) (114 051 SMILES), and NORMAN Substance Database (NORMAN SusDat) (106 632 SMILES). Additionally, we considered 42 648 SMILES from the NIST20, Agilent METLIN Metabolomics, MSDial, and GNPS databases. After removing duplicates, the initial database consisted of 2 211 691 unique SMILES.

To be as exhaustive as possible in the annotation of unidentified spectra from ARUS, we enlarged the *in silico* database with all compounds available for download from PubChem, resulting in an extended database of 96 492 904 predicted spectra.

To ensure consistency across the SMILES obtained from different databases, we calculated their canonical SMILES using the Rdkit package (https://www.rdkit.org). This process allowed us to identify how many SMILES corresponded to the same compound.

Key Points

- We tackle the problem of predicting the fragmentation spectrum of metabolites and small molecules.
- Whereas existing *in silico* fragmentation tools based on deep learning typically use a single model to predict the whole fragmentation spectrum, we propose to model each individual fragment separately.
- We achieve more accurate spectral predictions than the state of the art in both rule-based methods and deep learning methods.
- SingleFrag is fast enough to allow us to create a database with millions of predicted spectra, and use this database to annotate metabolites.
- Besides validating this annotation method on a test set with ground truth, we apply it to recurrent unidentified spectra from the ARUS database, successfully annotating three new metabolites.

Acknowledgements

This research was supported by projects PID2022-142600NB-I00 (MS-P and RGu), PID2022-136226OB-I00 (OY), and FPI grant PRE2020-095552 (OCT), from MCIN/AEI/10.13039/501100011033 and ERDF/EU, by projects 2021SGR-633 (MS-P. and RGu) and 2021SGR842 (OY) from the Government of Catalonia, and by the European Union NextGenerationEU/PRTR. MFK has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945413 and from Universitat Rovira i Virgili.

Author contributions

M.S.-P. and R.Gu. designed the research. M.P.-R. and R.Gu. wrote code. M.P.-R. performed experiments. M.P.-R., M.F.K., R.Gi. and J.M.B. collected and processed spectral data. S.J. performed the manual elucidation of ARUS spectra. M.P.-R., O.Y., M.S.-P., and R.Gu. analyzed and discussed results. M.P.-R., R.Gi., O.Y., M.S.-P., and R.Gu. wrote the paper.

Conflict of interest: The authors declare no conflict of interest.

Data availability

A Python implementation of SingleFrag and trained SingleFrag models are available from https://github.com/MaribelPR/SingleFrag.

References

1. Beger R. *et al.* Metabolomics enables precision medicine: "a white paper, community perspective". *Metabolomics* 2016;**12**:149. https://doi.org/10.1007/s11306-016-1094-6
2. Qiu S, Cai Y, Yao H. *et al.* Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Sig Transduct Target Ther* 2023;**8**:132. https://doi.org/10.1038/s41392-023-01399-3
3. Rafiq T, Azab SM, Teo KK. *et al.* Nutritional metabolomics and the classification of dietary biomarker candidates: a critical review. *Adv Nutr* 2021;**12**:2333–57. https://doi.org/10.1093/advances/nmab054
4. Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV. *et al.* Mass spectrometry-based metabolomics in microbiome investigations. *Nat Rev Microbiol* 2022;**20**:143–60. https://doi.org/10.1038/s41579-021-00621-9
5. Viant MR, Ebbels TMD, Beger RD. *et al.* Use cases, best practice and reporting standards for metabolomics in regulatory toxicology. *Nat Commun* 2019;**10**:3041. https://doi.org/10.1038/s41467-019-10900-y
6. Stein S. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal Chem* 2012;**84**:7274–82. https://doi.org/10.1021/ac301205z
7. Giera M, Yanes O, Siuzdak G. Metabolite discovery: Biochemistry's scientific driver. *Cell Metab* 2022;**34**:21–34. https://doi.org/10.1016/j.cmet.2021.11.005

8. Vinaixa M, Schymanski EL, Neumann S. *et al.* Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends Anal Chem* 2016;**78**:23–35. https://doi.org/10.1016/j.trac.2015.09.005
9. Frainay C, Schymanski EL, Neumann S. *et al.* Mind the gap: mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites* 2018;**8**:51. https://doi.org/10.3390/metabo8030051
10. Blaženović I, Kind T, Ji J. *et al.* Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* 2018;**8**:31. https://doi.org/10.3390/metabo8020031
11. Dührkop K, Fleischauer M, Ludwig M. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**:299–302. https://doi.org/10.1038/s41592-019-0344-8
12. Aguilar-Mogas A, Sales-Pardo M, Navarro M. *et al.* iMet: a network-based computational tool to assist in the annotation of metabolites from tandem mass spectra. *Anal Chem* 2017;**89**: 3474–82. https://doi.org/10.1021/acs.analchem.6b04512
13. Wang Y, Kora G, Bowen BP. *et al.* MIDAS: a database-searching algorithm for metabolite identification in metabolomics. *Anal Chem* 2014;**86**:9496–503. https://doi.org/10.1021/ac5014783
14. Ruttkies C, Schymanski EL, Wolf S. *et al.* MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Chem* 2016;**8**:3. https://doi.org/10.1186/s13321-016-0115-9
15. Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinf* 2019;**20**:376. https://doi.org/10.1186/s12859-019-2954-7
16. Wang F, Liigand J, Tian S. *et al.* CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem* 2021;**93**:11692–700. https://doi.org/10.1021/acs.analchem.1c01465
17. Wei JN, Belanger D, Adams RP. *et al.* Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Centr Sci* 2019;**5**:700–8. https://doi.org/10.1021/acscentsci.9b00085
18. Young, A., Röst, H. & Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence* **6**, (2024). Publisher: Nature Publishing Group, 404, 16. https://doi.org/10.1038/s42256-024-00816-8
19. Hong Y, Li S, Welch CJ. *et al.* 3DMolMS: prediction of tandem mass spectra from 3D molecular confo rmations. *Bioinformatics* 2023;**39**:btad354. https://doi.org/10.1093/bioinformatics/btad354
20. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 2018;**58**: 27–35. https://doi.org/10.1021/acs.jcim.7b00616
21. Zhou J, Cui G, Hu S. *et al.* Graph neural networks: a review of methods and applications. *AI Open* 2020;**1**:57–81. https://doi.org/10.1016/j.aiopen.2021.01.001
22. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. *Nat Mach Intell* 2021;**3**:1023–32. https://doi.org/10.1038/s42256-021-00418-8
23. Jiang D, Wu Z, Hsieh CY. *et al.* Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Chem* 2021;**13**:12. https://doi.org/10.1186/s13321-020-00479-8
24. Simón-Manso Y, Marupaka R, Yan X. *et al.* Mass spectrometry fingerprints of small-molecule metabolites in biofluids: building a spectral library of recurrent spectra for urine analysis. *Anal Chem* 2019;**91**:12021–9. https://doi.org/10.1021/acs.analchem.9b02977
25. SeesLab, G. Created in biorender. https://BioRender.com/z67l779 (2025). Figure created with BioRender.com.
26. Wishart DS, Guo AC, Oler E. *et al.* HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2021;**50**:D622–31. https://doi.org/10.1093/nar/gkab1062
27. Horai H, Arita M, Kanaya S. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 2010;**45**:703–14. https://doi.org/10.1002/jms.1777
28. Elapavalore A, Kondić T, Singh RR. *et al.* Adding open spectral data to massbank and pubchem using open source tools to support non-targeted exposomics of mixtures. *Environ Sci: Processes Impacts* 2023;**25**:1788–801. https://doi.org/10.1039/d3em00181d
29. Sawada Y, Nakabayashi R, Yamada Y. *et al.* Riken tandem mass spectral database (respect) for phytochemicals: a plant-specific ms/ms-based data resource and database. *Phytochemistry* 2012;**82**:38–45. https://doi.org/10.1016/j.phytochem.2012.07.007
30. Xing S, Shen S A, Xu B. *et al.* BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nat Methods* 2023;**20**: 881–90. https://doi.org/10.1038/s41592-023-01850-x
31. Tanimoto TT. *Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation, 1958.

Briefings in Bioinformatics, 2025, **26(4)**, bbaf333
https://doi.org/10.1093/bib/bbaf333
Case Study