FIDDLE: a deep learning method for chemical formulas prediction from tandem mass spectra

Yuhui Hong[1], Sujun Li[1], Yuzhen Ye[1], Haixu Tang[1*]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, 700 N. Woodlawn Avenue, Bloomington, 47408, Indiana, USA.

*Corresponding author(s). E-mail(s): hatang@iu.edu;
Contributing authors: yuhhong@iu.edu; sujli@indiana.edu; yye@iu.edu;

Abstract

Molecular identification through tandem mass spectrometry is fundamental in metabolomics, with formula identification serving as an initial step in the process. Current computational methods often struggle with accuracy, speed, and scalability for relatively larger molecules, limiting high-throughput workflows. We present FIDDLE (**F**ormula **ID**entification by **D**eep **LE**arning), a deep learning-based method trained on over 38,000 molecules and 1 million MS/MS spectra from various Quadrupole Time-of-Flight (QTOF) and Orbitrap instruments. FIDDLE accelerates formula identification by more than 10-fold and achieves top-1 and top-5 accuracies of 88.3% and 93.6%, respectively, outperforming state-of-the-art methods based on top-down (SIRIUS) and bottom-up (BUDDY) approaches by over 10%. On external metabolomics datasets, FIDDLE achieves top-3 accuracies of 72.1% (positive ion mode) and 66.2% (negative ion mode), with further improvements to 80.0% and 73.8% when combined with SIRIUS and BUDDY.

1 Introduction

Tandem mass spectrometry (MS/MS) is an essential analytical tool for identifying small molecules and elucidating their structural characteristics. The standard approach for identifying unknown analytes from MS/MS data involves searching spectra against reference spectral libraries [1–3]. However, due to limitations in time,

labor, and resources, a significant portion of chemical signatures remains uncharacterized, often termed "dark matter" [4]. These unidentified small molecules may possess unique bioactivities and play crucial roles in understanding biological mechanisms. Unfortunately, such molecules may lack corresponding reference spectra in spectral libraries or may not have been previously reported in the literature (i.e., the 'unknown unknowns' [5]). As a result, the identification of unknown compounds has become a challenging yet vital research area, spanning metabolomics, environmental analysis, natural product and drug discovery, and more [6–11]. Specifically, the prediction of molecular formulas serves as the initial and most fundamental step, providing critical constraints that facilitate structural elucidation and the annotation of fragments for these unknown compounds [12, 13].

While MZmine initially incorporated isotope pattern matching, MS/MS fragmentation analysis, and heuristic rules as a toolbox [14], computational methodologies for chemical formula identification from MS/MS data have evolved into top-down and bottom-up approaches, exemplified by SIRIUS [15] and BUDDY [16], respectively. SIRIUS begins by generating candidate formulas through the analysis of isotope patterns [17], then computes a fragmentation tree for each candidate [18] and evaluates them by comparing theoretical fragmentation patterns against experimental MS/MS data. The evaluation metric considers multiple factors including fragment masses, intensities, and isotope patterns to estimate the likelihood of each candidate producing the observed spectrum. However, SIRIUS's reliance on neutral loss fragments for tree generation poses a limitation, as it does not account for fragments in multiply-charged spectra that may include charged lost fragments, thus limiting its ability to handle multiply-charged spectra. Additionally, its efficiency is hindered by the computational demand of generating fragmentation trees for all potential formulas inferred from isotope patterns. MIST-CF shows that fragmentation trees can be replaced with a simple peak subformula assignment routine, achieving equally accurate and fast predictions [19]. However, it still relies on SIRIUS's algorithmic decomposition of exact masses into formula candidates limiting efficiency and accuracy. In contrast, BUDDY significantly reduces the number of candidate formulas by focusing on those explainable by MS/MS data, using a reference library of known formulas. It ranks candidates matching the precursor mass and estimates a false discovery rate to provide a confidence score. However, BUDDY's scope is restricted by the coverage of its reference library, potentially missing entirely uncharacterized and previously unreported formulas. Our experiments revealed that more than 300 unique formulas represented in MS/MS spectra in the National Institute of Standards and Technology (NIST) Spectral Library, Massbank of North America (MoNA) [2], Global Natural Product Social Molecular Networking (GNPS) Spectral Library [20], and the Agilent Personal Compound Database and Library (PCDL) fall outside the MS/MS-explainable space of BUDDY, rendering them unanalyzable by the method.

Both computational methods underutilize the full scope of information present in MS/MS spectrum data. SIRIUS 6, for instance, considers a limited number of peaks (up to 60), while BUDDY relies on manually extracted features, such as double-bond equivalent values of annotated fragments. As a result, as the precursor mass-to-charge ratio (m/z) increases, the computational complexity also increases, but their accuracy

tends to decrease significantly due to the increasing number of candidate molecular formulas that can match the precursor mass, which expands the search space and raises the potential for ambiguity. This limitation stems from the fact that higher precursor m/z values often correspond to larger, more complex molecular structures, when then requires the evaluation of a broader range of potential formulas. Furthermore, the peaks excluded from analysis, along with the relationships between considered and unconsidered peaks, may hold crucial structural information that is not exploited.

In this paper, we address these limitations by introducing a deep learning approach to chemical formula identification. We present FIDDLE (**F**ormula **ID**entification from tandem mass spectra by **D**eep **LE**arning), which employs dilated convolutions with large kernels [21, 22] to extract high-dimensional representations of MS/MS data and predict candidate formulas. These initial predictions are refined using a breadth-first search algorithm that adjusts atomic compositions to align the candidate formulas with the precursor mass. Additionally, we train a secondary deep learning model to estimate confidence scores for candidate formulas and rank them based on MS/MS features learned from the formula identification model. Compared to traditional computational methods, our approach dramatically reduces the candidate formula space for a given MS/MS spectrum to a small number (at most 5 formulas by default). This reduction is facilitated by the deep learning model, which benefits from accelerated GPU-based tensor computations. The narrowed candidate space simplifies confidence score estimation, and the MS/MS features learned by the deep learning model can be reused for this purpose. The source code for all experiments is publicly available at https://github.com/JosieHong/FIDDLE.

2 Results

2.1 Deep learning method for formula identification

Predicting target formulas directly from MS/MS spectra under varying experimental conditions presents significant challenges. To address this, we break down the task into three steps as illustrated in Fig. 1 (a): 1) predicting formulas from MS/MS spectra using a deep learning model; 2) generating candidate formulas using a breadth-first formula refinement algorithm; and 3) calculating confidence scores for the candidate formulas using an additional deep learning model. The formula refinement step relaxes the requirement for exceptionally high initial prediction accuracy by allowing adjustments to candidate formulas with minimal atom modifications. Moreover, since assessing the correctness of a limited set of predictions is easier than identifying the top correct outcomes from an infinitely large pool [23], the refinement step also improves the confidence score estimation.

To input an MS/MS spectrum into a deep learning model, we first bin it into a 1-D vector with a fixed mass-to-charge ratio (m/z) resolution. For example, an MS/MS spectrum with a maximum m/z of 1500 Da is binned into a vector of length 7500, with each bin representing a resolution of 0.2 Da. Molecular formulas are directly converted into formula vectors, where each element type is represented by its atom count as the corresponding value in the vector. For instance, the molecular formula $C_6H_{12}O_6$ can be represented as the vector $[6, 12, 6, 0, ...]$, where the first three integers correspond

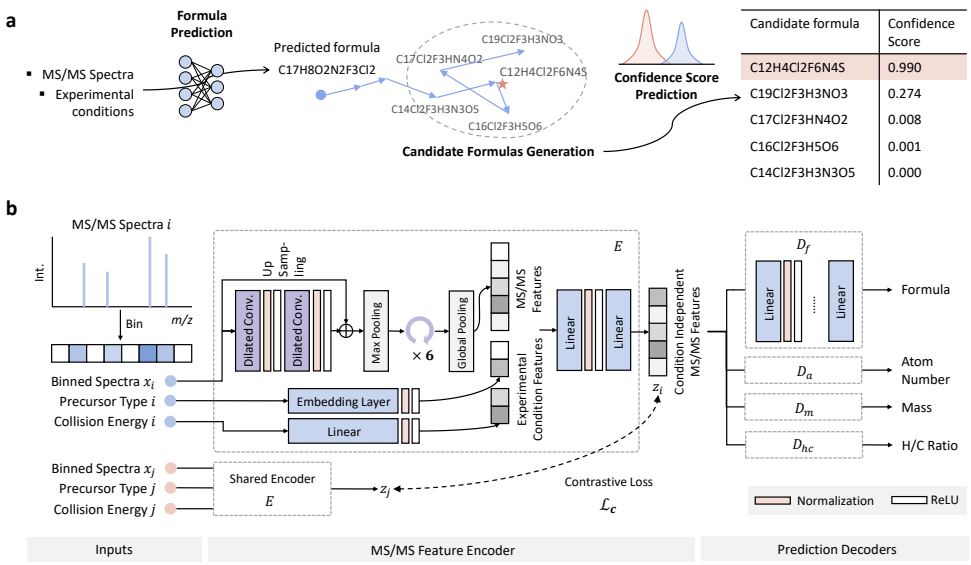


Fig. 1: **formula identification from tandem mass spectra using deep learning.** (a) The FIDDLE workflow comprises three main steps: predicting formulas from MS/MS spectra using a deep learning model; generating candidate formulas through a breadth-first formula refinement algorithm; and predicting confidence scores for the candidate formulas. (b) The deep learning model architecture includes an MS/MS spectrum encoder (E) and decoders (D_f, D_a, D_m, and D_{hc}) that outputs the predicted formula along with auxiliary variables, such as atom count, molecular mass, and H/C ratio.

to the number of carbon, hydrogen, and oxygen atoms, respectively, followed by atom counts for other elements in a predefined order.

To encode the MS/MS spectra, we use stacked dilated convolutions with large kernels to capture relationships between peaks across broad mass ranges [21, 22]. The learned MS/MS features are then concatenated with experimental conditions, such as collision energies, precursor types, and experimental precursor m/z, and fed into two linear layers to produce condition-independent MS/MS features, denoted as $\mathbf{z_i}$ and $\mathbf{z_j}$ in Fig. 1 (b). During training, a contrastive loss is applied to ensure that these condition-independent MS/MS features are close for spectra from the same molecule and far apart for spectra from different molecules. We use sequential linear layers as decoders for both formula identification and auxiliary tasks, including atom number prediction, molecular mass prediction, and H/C ratio prediction. These auxiliary tasks, incorporated through multitask learning, enhance the generalizability of the deep learning model and serve as a form of regularization [24]. The parameters of the model are specified in Section 4.2.

Recognizing that deep learning models cannot guarantee the validity or perfect accuracy of predicted formulas, we developed a breadth-first formula refinement algorithm. This algorithm aims to make minimal adjustments to atom counts to ensure that the formulas comply with SENIOR rules [25] and align with the target mass within a specified mass tolerance—specifically, 10 parts per million (ppm) for MS/MS data from QTOF instruments and 5 ppm for data from Orbitrap instruments. This refinement process produces a set of k candidate formulas for each MS/MS spectrum, where k is set to 5 by default. Note that the algorithm is flexible enough to integrate results from various formula identification methods, such as SIRIUS and BUDDY, where the predicted formula can be expanded into a longer list of candidate formulas. An auxiliary model is then developed to estimate confidence scores using the MS/MS features learned during formula identification and each candidate formula. Finally, the candidate formulas are ranked based on their estimated confidence scores.

2.2 Performance of formula identification

MS/MS spectra were collected from NIST23, NIST20, Agilent PCDL, MoNA, and GNPS, as well as from an internal dataset (see Section 4.1.1 for details) acquired using a Waters QTOF mass spectrometer. The MS/MS spectra were preprocessed according to the methods described in Section 4.1, including filtering based on peak count, molecular mass, atom type and number, and mass difference in ppm. Additional preprocessing steps included simulating precursor m/z values for the NIST dataset, simplifying precursor types, and constructing the training set for contrastive learning. In total, 131,224 MS/MS spectra from 15,399 molecules acquired with QTOF mass spectrometers and 965,656 MS/MS spectra from 28,383 molecules acquired with Orbitrap mass spectrometers were used for training and evaluation. A summary of the number of spectra and compounds in each dataset is provided in Table 1. We retained MS/MS spectra from compounds found exclusively in NIST23 and not in any other libraries (including NIST20) as the test set, ensuring these spectra were not used during the training of any of the models being compared. It is important to note that these spectra were published after the release of BUDDY and SIRIUS. Hence, it is important not to leverage them to train FIDDLE and maintain a fair comparison. Consistent with previous studies [16], we used the *top-k accuracy* to evaluate the performance of formula identification algorithms. This metric is calculated as the proportion of spectra for which the correct formula is included among the top k (by default $k = 5$) ranked formulas predicted by a given algorithm. The settings for comparison methods are specified in Section 4.6.

As shown in Fig. 2 (a)-(f), FIDDLE outperforms the other state-of-the-art formula identification algorithms (BUDDY and SIRIUS), and speeds up the cumulative runtime by approximately 10-fold compared to BUDDY and 100-fold compared to SIRIUS. In addition to formula identification, performance metrics for auxiliary tasks, including mass, atom number, and H/C ratio predictions, are presented in Extended Fig. 1. Notably, the accuracy of BUDDY and SIRIUS declines significantly for larger compounds (molecular weight > 800 Da; Fig. 2 (b) and (e)) compared to smaller ones (Fig. 2 (a) and (d)). In contrast, FIDDLE achieves top-1 accuracies of 0.642 and 0.813 for QTOF and Orbitrap spectra, respectively. For both small and large compounds,

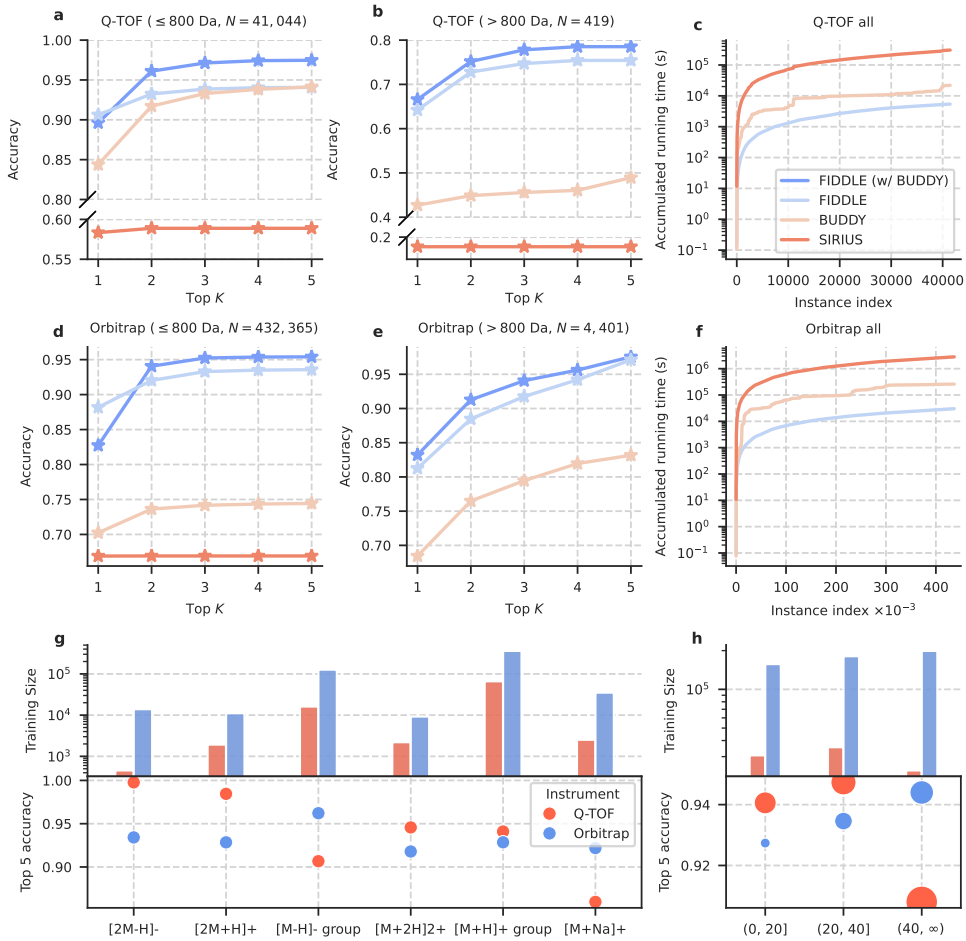


Fig. 2: **Comparison of the performance of FIDDLE with SIRIUS and BUDDY on formula identification.** (a) and (b) show top-k accuracy formula identification for QTOF MS/MS spectra of molecules with masses $\leq$ 800 Da and $>$ 800 Da, respectively. N represents the number of spectra. (c) and (f) show cumulative running times for SIRIUS, BUDDY, and FIDDLE across spectra sorted by molecular mass. (d) and (e) present the same accuracy measurements as (a) and (b) for Orbitrap MS/MS spectra. Notably, SIRIUS did not produce any output for the Orbitrap spectra of large molecules, likely because the running time exceeded the preset maximum limit. Consequently, these results are not included in (e). FIDDLE supports all 14 precursor types with charges of +1, -1, and +2, while BUDDY and SIRIUS are evaluated only on their supported types, excluding compounds that trigger the timeout. (g) and (f) illustrate the impact of precursor types and collision energies on the overall performance.

FIDDLE reaches accuracies of 0.941 and 0.754 for QTOF spectra and 0.936 and 0.971 for Orbitrap spectra when considering the top-5 candidate formulas. Moreover, incorporating BUDDY's results into FIDDLE's candidate formula pool further improved prediction accuracy, indicating that BUDDY and FIDDLE perform well on different sets of compounds and can be complementary. While combining these methods may require additional running time, it can yield superior overall results.

We integrated precursor type and collision energy as metadata in the deep learning model and evaluated performance under different conditions, as shown in Fig. 2 (g) and (h). FIDDLE demonstrates optimal performance when analyzing spectra of dimers $[2\,\mathrm{M}+\mathrm{H}]^+$ and $[2\,\mathrm{M}-\mathrm{H}]^-$, perhaps due to their more predictable fragmentation patterns and simpler spectral signatures (Extended Fig. 2). Furthermore, higher collision energies produce more peaks, enhancing pattern recognition and boosting performance. However, only the subset of spectra with collision energies in the range $[40, \infty)$ for the QTOF instrument had sufficient training data for optimal results by FIDDLE.

2.3 Ablation study

The MoNA dataset is used to illustrate the improvements of FIDDLE achieved through data augmentation, contrastive loss, and post-processing steps, including candidate formula generation and confidence score prediction. We evaluated FIDDLE's performance across different processing steps by measuring the proportion of top-1 formulas with varying numbers of missed atoms (Fig. 3 (a)) and missed heavy atoms (Fig. 3 (c)), respectively. Because hydrogen (approximately 1.008 Da) is light and difficult to accurately determine, the count of heavy atoms (excluding hydrogen) is also considered. The number of missed atoms is calculated by summing the differences in atom counts across all elements (e.g., carbon, oxygen, nitrogen). A comparison of FIDDLE's performance in 'Pred w/o DA' (prediction without data augmentation) and 'Pred' (prediction with data augmentation) shows that data augmentation significantly increases the proportion of correctly predicted formulas (from 0% to 9% for all atoms, and from 18% to 26% for heavy atoms). The contrastive loss further enhances performance, as seen in results labeled 'Pred w/o CL'. Comparing the performance of FIDDLE under 'Pred' with 'Post Top-1' (top-1 accuracy after post-processing) to 'Post Top-5' (top-5 accuracy after post-processing) shows that FIDDLE's candidate formulas cover the correct formulas for more than 84% of the MS/MS spectra. These candidates are subsequently ranked based on their confidence scores predicted by FIDDLE, achieving an AUC (area under the ROC curve) of 0.97, as shown in Fig. 3 (b). From Fig. 3 (d), it is clear that correct and incorrect formulas can be effectively distinguished based on their confidence scores. Notably, after post-processing, no formulas with three or fewer missed atoms remain, indicating that while post-processing may occasionally worsen certain incorrect predictions, most incorrect formulas with only a small number of missing atoms are effectively refined.

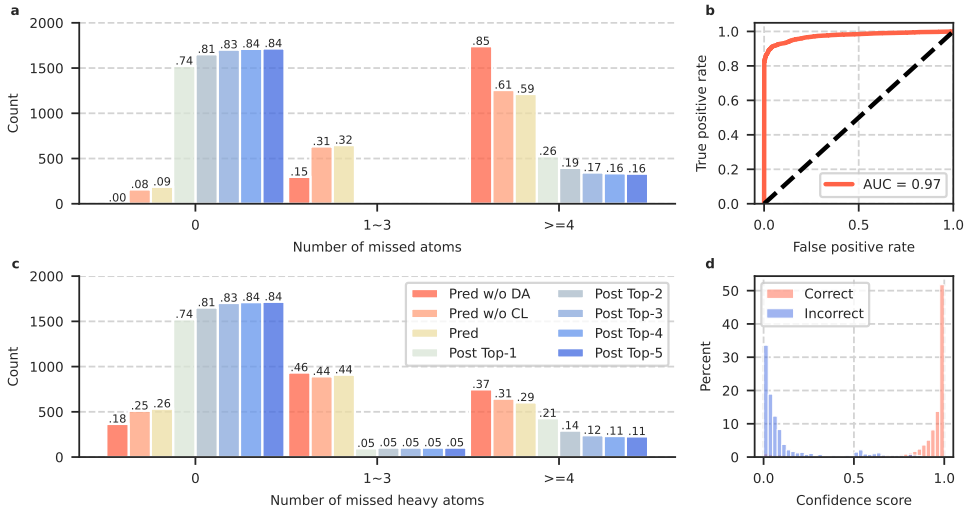


Fig. 3: **Ablation study of data augmentation, contrastive loss, and post-processing bin FIDDLE.** (a) and (c) show the proportion of top-1 formulas with different numbers of missed atoms (a) and missed heavy atoms (c), respectively, predicted by FIDDLE with different processing steps, where 'Pred w/o DA' indicates the model trained without data augmentation; 'Pred w/o CL' indicates the model trained without contrastive learning; 'Pred' indicates the model trained with contrastive loss; 'Post Top-1' to 'Post Top-4' indicate the top-1 to top-4 formulas from post-processing. (b) and (d) show the AUC-ROC curve (b) and the distributions (d) for discriminating correctly and incorrectly predicted formulas based on their predicted confidence scores among the top-5 predicted formulas from the MS/MS spectra in the MoNA dataset.

2.4 Evaluation using benchmarking metabolite datasets

Next, we compared the performance of FIDDLE against SIRUIS, BUDDY, and MIST-CF on three benchmarking metabolite datasets, the Critical Assessment of Small Molecule Identification (CASMI) 2016 [26], CASMI 2017, and EMBL-MCF 2.0 [27] datasets, respectively. For a fair comparison, we removed overlapping compounds in these datasets with our training set; in total 181 (231 spectra), 2 (2 spectra), and 107 (184 spectra) were removed from these three datasets, respectively (for details see Table 2). To compare with MIST-CF, we train MIST-CF using the same training and test data as used for FIDDLE*, excluding unsupported precursor types from both datasets for MIST-CF (see details in Extended Table 4).

As illustrated in Fig. 4, FIDDLE performs comparably well with SIRIUS and BUDDY on the two CASMI datasets, while demonstrating significantly better performance on the EMBL-MCF 2.0 dataset. Since all methods achieve near-optimal performance on top-3 accuracy, top-4 and top-5 accuracies are not demonstrated here. The slightly worse performance of FIDDLE on CASMI can be attributed to multiple factors, including lower MS/MS spectral quality (CASMI 2016 and 2017 were collected

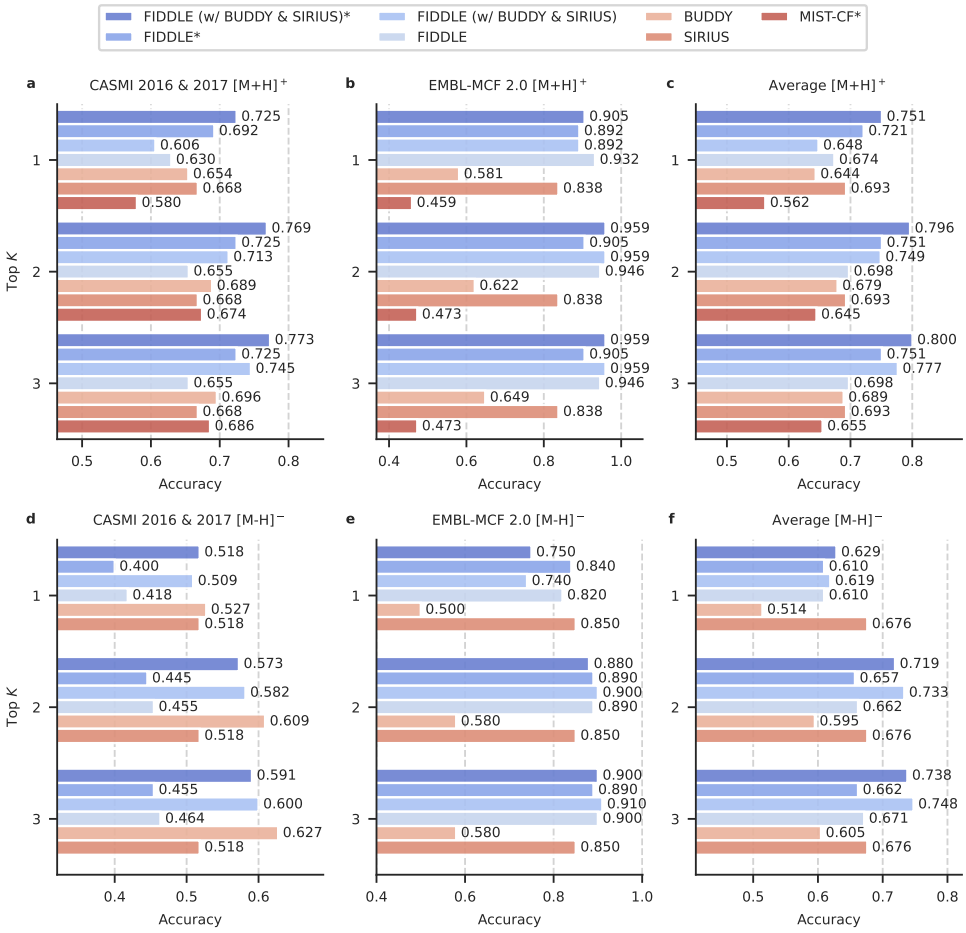


Fig. 4: **Performance of formula identification on external test sets.** Methods marked with an asterisk (*) denote deep learning models trained on all available datasets, whereas models without an asterisk were trained on all available datasets except for the NIST23 dataset. Following the original configuration of MIST-CF, we only evaluated it in positive ion mode.

on a QTOF instrument), lower compound similarity to the training set (see Extended Fig. 3), among others. The computation-based methods exhibit greater sensitivity to mass deviation, as their performance decreases on the EMBL-MCF 2.0 dataset with larger mass deviations compared to the CASMI dataset, as shown in Extended Fig. 3 (b). Notably, incorporating the NIST23 dataset into the training set further enhances FIDDLE's performance on these external test sets, resulting in higher or comparable top-3 accuracies compared to SIRIUS, BUDDY, and MIST-CF. Combining candidate formulas from all methods and ranking them by FIDDLE's predicted confidence scores improves performance. On average, across three test sets, our method achieves top-3

accuracies of 80.0% (positive) and 73.8% (negative) ion mode, while BUDDY, SIRIUS, and MIST-CF achieve average top-3 accuracies of 68.9% (positive) and 60.5% (negative), 69.3% (positive) and 67.6% (negative), and 65.5% (positive), respectively.

3 Discussion

In this work, we introduce a deep learning approach named FIDDLE to identify chemical formulas from tandem mass spectra. It consists of three steps: predicting formulas from tandem mass spectra, generating candidate formulas, and ranking the candidates based on predicted confidence scores. FIDDLE not only accelerates the formula identification compared to state-of-the-art algorithms (BUDDY and SIRUIS), but also outperforms them on both the evaluation set (with 88.3% top-1 accuracy and 93.6% top-5 accuracy) and external benchmarking data sets (with the average top-3 accuracy of 80.0% and 73.8% for positive and negative ion modes, respectively, across three datasets).

According to our ablation study, both data augmentation and contrastive loss between manually combined spectra pairs enhance formula identification from tandem mass spectra. The post-processing steps, which include generating candidate formulas and ranking them by predicted confidence scores, aim to refine the predicted formulas using the SENIOR rules. These steps refine most candidate formulas with no more than three missed heavy atoms, and significantly alleviate the challenge of incorrect hydrogen counts due to their small mass. Furthermore, because FIDDLE and other formula identification methods perform well on different compounds and spectra, combining their predicted candidate formulas and ranking them by using FIDDLE's confidence score prediction further improves the accuracy of the predicted formulas, even though its running time may be longer.

We note that, even though the contrastive loss is employed to take into account the effect of experimental conditions, this effect is not completely eliminated, as FIDDLE still shows variations in accuracy under different conditions. Future work may be focused on improving accuracy across different conditions, especially those with less training data. The workflow of FIDDLE offers several avenues, such as improving MS/MS representations by pre-trained models from self-supervise learning [28], using MS/MS prediction methods for data augmentation [29], and extending the workflow with adduct type prediction for comprehensive application.

Beyond formula identification, FIDDLE serves as a robust foundation for compound structural elucidation from MS/MS spectra, enabling the inference of covalent bonds between atoms based on their atomic composition. In the future, this methodology could be expanded to characterize diverse molecular structures by leveraging machine learning techniques to extract structural information from MS/MS spectra and integrating this information into molecular structure characterization.

	QTOF		Orbitrap	
	# Spec	# Mol	# Spec	# Mol
NIST23	32,057	2,490	569,157	27,836
NIST20	31,749	2,479	394,395	17,883
Agilent PCDL	42,410	11,479	0	0
MoNA	21,609	2,652	591	303
GNPS	2,642	1,270	1,513	509
Waters QTOF	757	610	0	0
Unique	131,224	15,399	965,656	28,838

Table 1: Total numbers of spectra and compounds used for training and evaluation.

4 Methods

4.1 MS/MS data pre-processing

4.1.1 MS/MS data filtering

The training and testing MS/MS datasets are compiled from several sources, including NIST20, NIST23, MoNA, GNPS, Agilent PCDL, and an internal dataset, following the preprocessing methodology described by Hong et al. [29]. To construct the internal dataset, we acquired 1,424 compounds from APExBIO [30] and measured their tandem mass spectra using a Waters Synapt G2 mass spectrometer at 40 eV with various precursor types, including $[\mathrm{M}+\mathrm{H}]^+$, $[\mathrm{M}+\mathrm{Na}]^+$, $[2\,\mathrm{M}+\mathrm{H}]^+$, $[\mathrm{M}+2\,\mathrm{H}]^{2+}$, and $[\mathrm{M}-\mathrm{H}]^-$. To minimize the interference from impurities in the air, e.g., water vapor (H_2O) and carbon dioxide (CO_2), we excluded peaks below 50 m/z from the scans. Additionally, three benchmarking sets generated for community-wide evaluation of metabolite identification algorithms are used, including CASMI 2016, CASMI 2017, and EMBL-MCF 2.0.

These datasets undergo a series of filtering steps to ensure data quality: 1) Mass spectra with fewer than five peaks are excluded due to potential unreliability. 2) The m/z range is confined to $(0, 1500]$ to account for the rarity of spectra with m/z values above 1500. 3) Only the molecules composed of frequent atoms (C, H, O, N, F, S, Cl, P, B, I, Br, Na, and K) are retained. (iv) Only spectra associated with common precursor types are included, e.g. $[\mathrm{M}+\mathrm{H}]^+$, $[\mathrm{M}+\mathrm{Na}]^+$, $[2\,\mathrm{M}+\mathrm{H}]^+$, $[\mathrm{M}+\mathrm{H}-\mathrm{H_2O}]^+$, $[\mathrm{M}+\mathrm{H}-2\,\mathrm{H_2O}]^+$, $[\mathrm{M}+\mathrm{H}-\mathrm{NH_3}]^+$, $[\mathrm{M}+\mathrm{H}+\mathrm{NH_3}]^+$, $[\mathrm{M}+\mathrm{H}-\mathrm{CH_2O_2}]^+$, $[\mathrm{M}+\mathrm{H}-\mathrm{CH_4O_2}]^+$ for positive modes, and $[\mathrm{M}-\mathrm{H}]^-$, $[2\,\mathrm{M}-\mathrm{H}]^-$, $[\mathrm{M}-\mathrm{H}-\mathrm{CO_2}]^-$, $[\mathrm{M}-\mathrm{H}-\mathrm{H_2O}]^-$ for negative modes, along with $[\mathrm{M}+2\,\mathrm{H}]^{2+}$, a doubly charged precursor type. 4) The total count of atoms in the molecules is capped at 300 to exclude the molecules not typically classified as 'small molecules'. 5) The tolerance for precursor mass discrepancy is set at 10 ppm for QTOF and 5 ppm for Orbitrap instruments, ensuring precise mass matching. The statistics information of the datasets are detailed in the following Table 1 and Table 2. The specific instruments contained in each instrument type are shown in Extended Table 1.

	# Spec	# Mol	Instrument Type	Instrument
CASMI 2016	502	405	QTOF	Agilent 6540 QTOF and Waters Synapt G2 QTOF
CASMI 2017	40	40	QTOF	Waters Synapt G2 QTOF
EMBL-MCF 2.0	224	130	Orbitrap	Orbitrap Exploris 240 (ThermoFisher) and Q-Exactive Plus (ThermoFisher)

Table 2: Numbers of spectra and compounds from CASMI and EMBL-MCF 2.0 library.

4.1.2 Precursor m/z simulation

The precursor m/z from NIST20, NIST23, CASMI 2016, and CASMI 2017 are theoretical values so they are adjusted via random shifts, following the approach used by Xing et al. [16] and the observations by Böcker at el. [17] that the mass deviations fit a Gaussian distribution with the standard deviation of 1/3 of the mass tolerance. We sampled the deviations from Gaussian distributions within the set tolerance ranges (5 ppm for Orbitrap and 10 ppm for QTOF) to simulate the experimental conditions accurately. These simulated precursor m/z values are utilized throughout both the training and testing phases, enhancing the model's generalizability for application in real-world formula identification tasks.

4.1.3 Simplification of precursor types

As depicted in Fig. 2 (g), the dataset exhibits significant imbalance across different precursor types. To improve formula identification for less common precursor types, we simplify the precursor types by eliminating uncharged molecules, such as water (H_2O), ammonia (NH_3), carbon dioxide (CO_2), formic acid (CH_2O_2), and acetic acid (CH_4O_2), adding them into the original formula representation (for the detailed precursor types see Table Extended Table 2). For example, consider a molecular formula $C_6H_{12}N_2O_2$ with the precursor type $[M+H+NH_3]^+$. After simplifying the precursor type, the formula is adjusted to $C_6H_{15}N_3O_2$, and the precursor type is simplified to $[M+H]^+$, reflecting the integration of NH_3 into the molecular formula. This simplification is recorded allowing the predicted formulas to be converted back to the original formula with the original precursor type. Through this process, many uncommon precursor types are consolidated into more common precursor types, which expand the training data of common precursor types, thereby enhancing the predictions.

4.1.4 Training set construction and data augmentation

Each dataset, either acquired using QTOF or Orbitrap MS/MS instruments, is first split into training and test sets according to their molecular canonical SMILES, which ensures that there are no common compounds between the training set and the test set. Two strategies were employed for data splitting: 1) random splitting (see the model in Section 2.3 and the model marked with an asterisk in Section 2.4); 2) leaving spectra of unique compounds from NIST23 for evaluation (see the model in Section 2.2 and the model without an asterisk in Section 2.4). Then, for contrastive learning (CL), we

constructed MS/MS spectra pairs from the training sets. The spectra are grouped by canonical SMILES. For each compound, we randomly picked another compound from the same group to construct a positive pair and another compound from a different group to construct a negative pair.

To enhance the robustness of the deep learning model, we generated augmented MS/MS spectra by perturbing the spectra in the training set. Specifically, we added random noise sampled from a Gaussian distribution with a mean of 0 and a standard deviation of 0.1 ($\sim \mathcal{N}(0, 0.1)$) to the intensities of an experimental QTOF spectrum, generating two augmented spectra for each QTOF spectrum and thus tripling the size of the training set from QTOF mass spectrometry. On average, the cosine similarity between the augmented spectrum and the corresponding experimental spectrum is 0.936 , which is close to the similarity between of replicated spectra of the same compound (0.977) in the QTOF training set. However, noise adding lead to significantly lower similarity for Orbitrap spectra (as shown in Extended Fig. 4). Since the training size for Orbitrap is sufficiently large, we did not apply data augmentation for training deep learning model for Orbitrap.

4.2 Representation learning for MS/MS spectra

4.2.1 Model architecture

The representation learning for MS/MS spectra is structured as a two-stage process: MS/MS embedding and the elimination of the experimental condition effect. This approach decomposes the complex task of MS/MS representation learning under multiple experimental conditions into more tractable steps, facilitating the network's ability to extract meaningful features. Additionally, the resultant condition-independent MS/MS features reduce the complexity of the decoder.

Acknowledging the significance of the correlation among fragment ions in MS/MS, such as the neutral loss between two ions, dilated convolutions with large kernels are employed, building upon the previous work for *de novo* peptide sequencing [22]. Due to the large effective receptive field (ERF), this method allows the deep neural network to learn meaningful patterns across fragment ions with long mass ranges from the MS/MS data. Each convolutional block consists of two sequentialized dilated convolutions with ReLU activation, random dropout, a residual connection, and max-pooling. The max-pooling uses a kernel size of 2 and a stride size of 2, which halves the spectra length. Upsampling along the channel dimension is applied in the skip connection when the output and input channel numbers differ. Dilated convolutions with exponentially increasing dilation factors (1, 2, 4, 8, etc.) and large kernels rapidly expand the ERF without significantly increasing the number of parameters [21, 31]. We can calculate the ERF for our model as:

$$\text{ERF} = (\text{kernel size} - 1) \times (\text{dilation size} - 1) + 1 \tag{1}$$

where kernel sizes are 45, 43, 41, 39, 37, and 35, and the dilation sizes are 1, 2, 4, 8, 8, and 8, respectively. The total ERF of this model is 1153, indicating that each position in the MS/MS feature is directly influenced by 1153 input bins. It is worth

noting that while the ERF defines the range of direct input influence, the model can still capture patterns beyond this range. In addition, we utilize weight normalization to speed up the convergence and reduce the dependency of normalization on batch size, enabling training with small batches and limited GPU memory [32].

The features from all the blocks are concatenated together along the channel dimension, resulting in a 1024×235 tensor, where 1024 results from the stacked convolutional blocks with channel sizes of 32, 32, 64, 128, 256, and 512, and 235 is the length of the MS/MS spectra after pooling layers. Then, through a global pooling layer along the spectra-length dimension, the tensor is pooled into a 1024-dimensional vector as the embedded MS/MS feature. Precursor types, normalized collision energy, and simulated precursor m/z are included as experimental conditions. These conditions are embedded into a 16-dimensional vector through linear layers. The embedded MS/MS feature and experimental conditions are concatenated and fed into sequential linear layers for the condition-independent MS/MS features. These features are constrained by contrastive learning so that features from the same molecule are learned to be the same, while features from different molecules are learned to be different.

4.2.2 Loss function

Given a pair of binned spectra i and j, the encoder of the prediction model embeds them into condition-independent MS/MS feature vectors $\mathbf{z_i}$ and $\mathbf{z_j}$. These feature vectors are constrained by the contrastive loss:

$$\mathcal{L}_c(\mathbf{z_i}, \mathbf{z_j}) = (1 - y)||\mathbf{z_i} - \mathbf{z_j}||^2 + y \cdot \max(0, m - ||\mathbf{z_i} - \mathbf{z_j}||)^2 \tag{2}$$

where y is the label between the paired spectra. If spectra i and j are from the same molecule, the label is 1; otherwise, the label is 0. m is a hyperparameter (by default, $m = 1.0$) that defines the lower bound distance between spectra of different molecules.

Three auxiliary tasks, including the atom count prediction, the molecular mass prediction, and the prediction of the hydrogen-to-carbon (H/C) ratio, are used for model regularization to improve the model generalizability [24]. Mean squared error (MSE) losses are applied to these tasks. The loss function for sample i from the pair (i, j) is defined as:

$$\mathcal{L}_{t,i} = \alpha_t \cdot \mathrm{MSE}(y_{t,i}, \hat{y}_{t,i}) \tag{3}$$

where $\mathcal{L}_{t,i}$ denotes the loss for task t on sample i; $y_{t,i}$ and $\hat{y}_{t,i}$ denote the label and prediction, respectively; and α_t denotes the weight of loss on task t. In the experiments, we use the weights of 0.01, 1, and 10 for these three auxiliary tasks, respectively. Finally, the loss function for training the model is:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_{t,i} + \mathcal{L}_{t,j} + \mathrm{MSE}(\mathbf{f}_i, \hat{\mathbf{f}}_i) + \mathrm{MSE}(\mathbf{f}_j, \hat{\mathbf{f}}_j) \tag{4}$$

4.3 Formula refinement algorithm

The candidate formula refinement algorithm (Algorithm 1) refines an input chemical formula into a list of formulas that comply with the SENIOR rules [25] and match a target precursor mass within a specified tolerance. This algorithm can start with the

Algorithm 1 Candidate Formula Refinement Algorithm

Require:
Inputs: Initial formulas F_{init}, Target mass M
Settings: Tolerance ΔM, Maximum search depth $D_{\max}$, Number of formulas to return K
Ensure: $\Delta M > 0$
Initialization:
$F_{\text{refined}} \leftarrow []$ ▷ Stores refined formulas
$F_{\text{track}} \leftarrow []$ ▷ Tracks explored formulas
$F_{\text{cand}} \leftarrow F_{\text{init}}$ ▷ Initial candidates for refinement
$D_{\text{cand}} \leftarrow [0] \times \text{len}(F_{\text{init}})$ ▷ Search depths of candidates
Refinement Iteration:
while $\text{len}(F_{\text{refined}}) < K$ **and** F_{cand} **do**
  $f \leftarrow F_{\text{cand}}.\text{pop}()$, $d \leftarrow D_{\text{cand}}.\text{pop}()$, $F_{\text{track}}.\text{append}(f)$ ▷ Process next candidate
  if f passes SENIOR rules **and** mass of $f \in [M - \Delta M, M + \Delta M]$ **then**
    $F_{\text{refined}}.\text{append}(f)$
  end if
  if $d < D_{\max}$ **then** ▷ Explore new candidates
    Generate new candidates from f by editing one heavy atom
    $F_{\text{cand}}, D_{\text{cand}} \leftarrow$ Update with untracked new candidate, $d + 1$
  end if
  if Timeout check **then** ▷ Exit loop if timeout reached
    break
  end if
end while
return F_{refined} ▷ Returns refined formulas

formula predicted by the deep learning model or a formula identified by tools such as SIRIUS or BUDDY, facilitating the integration of results from multiple methods. The algorithm initializes lists to track refined formulas, previously explored formulas, and current candidates with their associated search depths. By iteratively modifying the counts of heavy atoms without exceeding a set search depth, the algorithm systematically refines each candidate formula. During the exploration, one heavy atom is added or removed, and hydrogen counts are adjusted to align with the precursor mass. The refinement process concludes either when a predefined number of formulas have been explored or when all candidate formulas are exhausted, incorporating a timeout check. The algorithm ultimately produces a list of formulas that align closely with the target precursor mass and adhere to the SENIOR rules.

4.4 Prediction of confidence score

The inputs for the confidence score prediction model consist of the concatenated condition-independent MS/MS feature vector and the vector representation of the predicted formula. The model architecture comprises a sequence of five linear layers, each followed by batch normalization, ReLU activation function, and random dropout

for regularization. The dimensions of the five layers are 416, 208, 104, 26, and 13, respectively. A Sigmoid activation function is used in the final layer to produce a confidence score between 0 and 1. Training data for this model is obtained from the inference results of initial formula identification followed by candidate formula refinement. Specifically, each of the top 5 refined candidate formulas is labeled, with 0 indicating a correct candidate and 1 denoting an incorrect one. Since this model has significantly fewer trainable parameters than the formula prediction model, the training set is randomly sampled down to 10,000 instances if it exceeds this limit to prevent overfitting.

4.5 Implementation and computing environment

The training of models in FIDDLE was conducted using a batch size of 512 over 200 epochs, with an initial learning rate of 0.001. The models were optimized using AdamW and a ReduceLROnPlateau learning rate schedule that reduces the learning rate when a metric has stopped improving for 5 epochs. The training device was two NVIDIA RTX A6000 GPUs equipped with 48 GB of memory. For inference only, the minimum GPU memory requirement is 500 MB when using a batch size of 1. The deep learning framework used was PyTorch version 1.13.0, running on Ubuntu 20.04.1.

4.6 Running SIRIUS and BUDDY

We installed SIRIUS v6.0.1, last updated on July 21st, 2024, for the Linux (64-bit) platform, accessible at https://github.com/boecker-lab/sirius. Additionally, the BUDDY v0.3.6 PyPI package, updated on January 29th, 2024, was obtained from https://github.com/Philipbear/msbuddy. All experiments were performed on a workstation with 20.04.1-Ubuntu 64-bit operating system equipped with 16 Intel(R) Core(TM) i7-9800X CPUs. For comprehensive configuration details, please refer to Extended Table 3.

5 Data availability

NIST20 and NIST23 libraries can be accessed with a license from NIST23 - Mass Spectral Library. Data from GNPS can be downloaded from GNPS - Mass Spectral Libraries. The MoNA, CASMI16, CAMI17, and EMBL-MCF 2.0 libraries can be downloaded from Mass Bank of North America. The Agilent PCDL is available from Agilent Technologies (https://www.agilent.com). The Waters QTOF dataset can be obtained by contacting us.

6 Code availability

The source code for all experiments is publicly available at GitHub: https://github.com/JosieHong/FIDDLE under the CC BY-NC-SA 4.0 License. The model trained on all the QTOF and Orbitrap MS/MS data can be downloaded from GitHub release: https://github.com/JosieHong/FIDDLE/releases/tag/v1.0.0.

Acknowledgements. We acknowledge the Center for Bioanalytical Metrology (CBM), an NSF Industry-University Cooperative Research Center, for providing funding under grant NSF IIP-1916645. This work was also partially supported by National Science Foundation grant DBI-2011271.

References

[1] 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).

[2] Horai, H. *et al.* Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry* **45**, 703–714 (2010).

[3] Blaženović, I. *et al.* Comprehensive comparison of in silico ms/ms fragmentation tools of the casmi contest: database boosting is needed to achieve 93% accuracy. *Journal of cheminformatics* **9**, 1–12 (2017).

[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] Stein, S. Mass spectral reference libraries: an ever-expanding resource for chemical identification (2012).

[6] Moco, S., Vervoort, J., Bino, R. J., De Vos, R. C. & Bino, R. Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry* **26**, 855–866 (2007).

[7] Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass spectrometry reviews* **26**, 51–78 (2007).

[8] Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature methods* **18**, 747–756 (2021).

[9] Hernandez, F., Sancho, J. V., Ibáñez, M. & Grimalt, S. Investigation of pesticide metabolites in food and water by lc-tof-ms. *TrAC Trends in Analytical Chemistry* **27**, 862–872 (2008).

[10] Dueñas, M. E. *et al.* Advances in high-throughput mass spectrometry in drug discovery. *EMBO Molecular Medicine* **15**, e14850 (2023).

[11] Liu, C. & Zhang, H. *High-Throughput Mass Spectrometry in Drug Discovery* (John Wiley & Sons, 2023).

[12] Wang, F. *et al.* Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analytical chemistry* **93**, 11692–11700 (2021).

[13] Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. Msnovelist: de novo structure generation from mass spectra. *Nature Methods* **19**, 865–870 (2022).

[14] Pluskal, T., Uehara, T. & Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, ms/ms fragmentation, heuristic rules, and isotope pattern matching. *Analytical chemistry* **84**, 4396–4403 (2012).

[15] Dührkop, K. *et al.* Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods* **16**, 299–302 (2019).

[16] Xing, S., Shen, S., Xu, B., Li, X. & Huan, T. Buddy: molecular formula discovery via bottom-up ms/ms interrogation. *Nature Methods* **20**, 881–890 (2023).

[17] Böcker, S., Letzel, M. C., Lipták, Z. & Pervukhin, A. Sirius: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**, 218–224 (2009).

[18] Rasche, F., Svatoš, A., Maddula, R. K., Böttcher, C. & Böcker, S. Computing fragmentation trees from tandem mass spectrometry data. *Analytical Chemistry* **83**, 1243–1251 (2011).

[19] Goldman, S., Xin, J., Provenzano, J. & Coley, C. W. Mist-cf: Chemical formula inference from tandem mass spectra. *Journal of Chemical Information and Modeling* **64**, 2421–2431 (2023).

[20] Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using gnps. *Nature protocols* **15**, 1954–1991 (2020).

[21] Lea, C., Vidal, R., Reiter, A. & Hager, G. D. Hua, G. & Jégou, H. (eds) *Temporal convolutional networks: A unified approach to action segmentation.* (eds Hua, G. & Jégou, H.) *Computer Vision – ECCV 2016 Workshops*, 47–54 (Springer International Publishing, Cham, 2016).

[22] Liu, K., Ye, Y., Li, S. & Tang, H. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications* **14**, 7974 (2023).

[23] Kwok, T.-Y. & Yeung, D.-Y. Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE transactions on neural networks* **8**, 630–645 (1997).

[24] Zhang, Y. & Yang, Q. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering* **34**, 5586–5609 (2021).

[25] Kind, T. & Fiehn, O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics* **8**, 1–20 (2007).

[26] Schymanski, E. L. *et al.* Critical assessment of small molecule identification 2016: automated methods. *Journal of cheminformatics* **9**, 1–21 (2017).

[27] Dekina, S., Alexandrov, T. & Drotleff, B. Embl-mcf 2.0: an lc-ms/ms method and corresponding library for high-confidence targeted and untargeted metabolomics using low-adsorption hilic chromatography. *Metabolomics* **20**, 114 (2024).

[28] Guo, H., Xue, K., Sun, H., Jiang, W. & Pu, S. Contrastive learning-based embedder for the representation of tandem mass spectra. *Analytical Chemistry* **95**, 7888–7896 (2023).

[29] Hong, Y. *et al.* 3dmolms: prediction of tandem mass spectra from 3d molecular conformations. *Bioinformatics* **39**, btad354 (2023).

[30] APExBIO. APExBIO: Achieve Perfection, Explore the Unknown. https://www.apexbt.com/. Accessed: 2023-04-04.

[31] Ding, X., Zhang, X., Han, J. & Ding, G. Dana, K., Hua, G., Roth, S., Samaras, D. & Singh, R. (eds) *Scaling up your kernels to 31×31: Revisiting large kernel design in cnns.* (eds Dana, K., Hua, G., Roth, S., Samaras, D. & Singh, R.) *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953–11965 (IEEE, 2022).

[32] Salimans, T. & Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems* **29** (2016).

7 Extended Data

Instrument Type	Instrument
QTOF	Agilent QTOF 6530 Agilent 6530 QTOF Waters Xevo G2 QTOF Agilent qTOF 6545 Agilent 6550 iFunnel
Orbitrap	Thermo Finnigan Elite Orbitrap Thermo Q Exactive HF Q Exactive Orbitrap (Thermo Scientific) Q Exactive Plus Orbitrap Thermo Scientific

Extended Table 1: List of QTOF and Orbitrap instruments.

Precursor Group	Precursor Type		Charge
$[M+H]^+$ group	$[M+H]^+$ $[M+H-2\,H_2O]^+$ $[M+H+NH_3]^+$ $[M+H-CH_4O_2]^+$	$[M+H-H_2O]^+$ $[M+H-NH_3]^+$ $[M+H-CH_2O_2]^+$ $[M+H-HCO]^+$	+1
$[2\,M+H]^+$	$[2\,M+H]^+$		+1
$[M+Na]^+$	$[M+Na]^+$		+1
$[M-H]^-$ group	$[M-H]^-$ $[M-H-H_2O]^-$	$[M-H-CO_2]^-$	-1
$[2\,M-H]^-$	$[2\,M-H]^-$		-1
$[M+2\,H]^{2+}$	$[M+2\,H]^{2+}$		+2

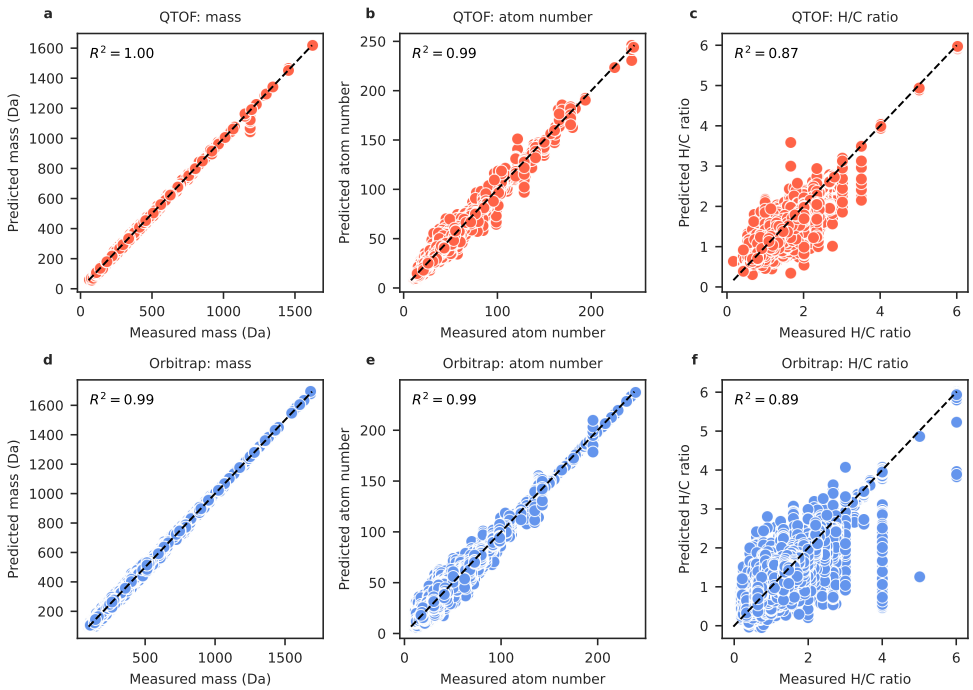
Extended Table 2: List of simplified precursor types.

	SIRIUS	BUDDY
Version	6.0.1	0.3.6
Atom Type Limitation	H, C, N, O, P, B, Cl, Br, S, F, I	H, C, N, O, P, B, Cl, Br, S, F, I
Precursor Type Limitation	$[M+H]^+$, $[M+H-H_2O]^+$, $[M+H-CH_4O_2]^+$, $[M+H-H_4O_2]^+$, $[M-H]^-$, $[M-H-H_2O]^-$	$[M+H]^+$, $[M+H-NH_3]^+$, $[M+H+NH_3]^+$, $[M+H+H_2O]^+$, $[M+H-H_2O]^+$, $[M+H-H_2O]^+$, $[M+H-HCO]^+$, $[M+K]^+$, $[M+Na]^+$, $[M-H]^-$, $[M-H+CH_2O_2]^-$, $[M-H+CH_4O_2]^-$, $[M+Cl]^-$, $[M+Br]^-$
Charge Limitation	+1, -1	+1, -1
Timeout for Each Compound	300 s	600 s
Mass Deviation Tolerance	10 ppm	10 ppm

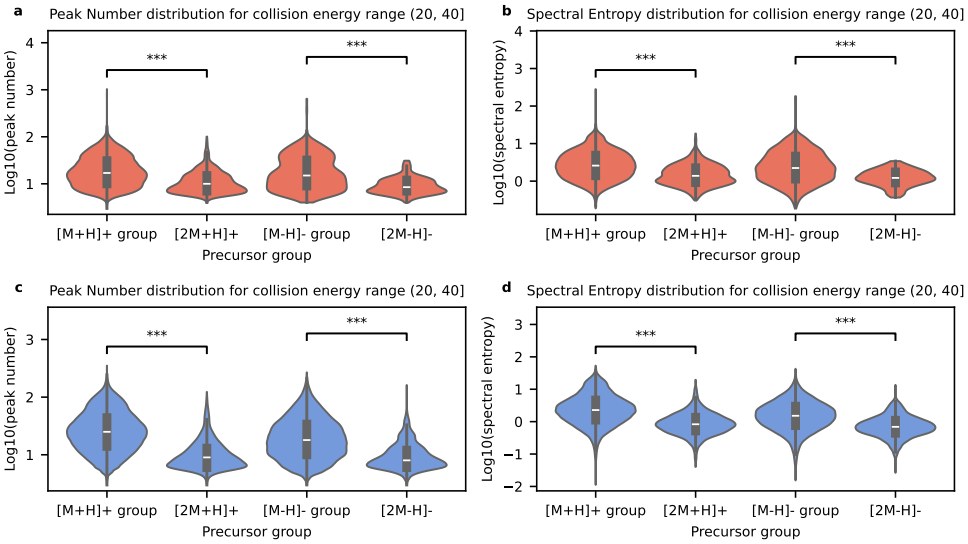
Extended Table 3: Configurations of SIRIUS and BUDDY.

Data Preprocessing	
Candidate Selection	"COMMON" filter [15] and FastFilter [19]
Atom Type Limitation	H, C, N, O, P, B, Cl, Br, S, F, I, Si, Se, Fe, Co, As, Na, K
Precursor Type Limitation	$[M+H]^+$, $[M+H-H_2O]^+$, $[M+H+NH_3]^+$, $[M+H-H_4O_2]^+$, $[M+Na]^+$, $[M+K]^+$
Charge Limitation	+1
Mass Deviation Tolerance	10 ppm
Maximum Number of Peaks for Subformula Assignment	20
Intensity Threshold of Peaks (Normalized to 1)	0.003
Model Architecture	
Number of Layers	2
Dropout Rate	0.1
Hidden Layer Dimensionality	128
Fourier Feature Representation	Absolute Sine Features
Training Parameters	
Batch Size	4
Maximum Number of Decoys	32
Maximum Number of Epochs	200
Learning Rate	0.00045
Learning Rate Decay Factor	0.88

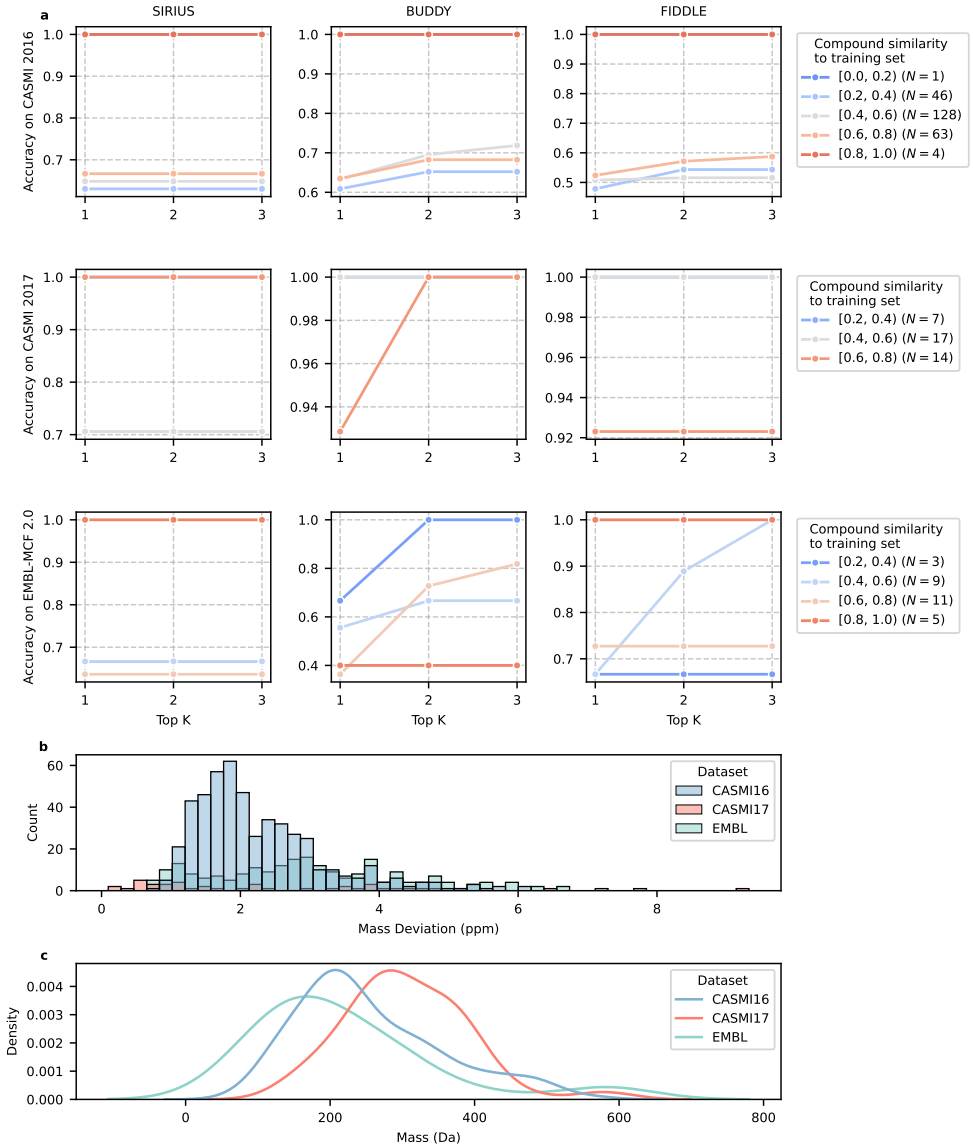
Extended Table 4: Configurations for training MIST-CF.



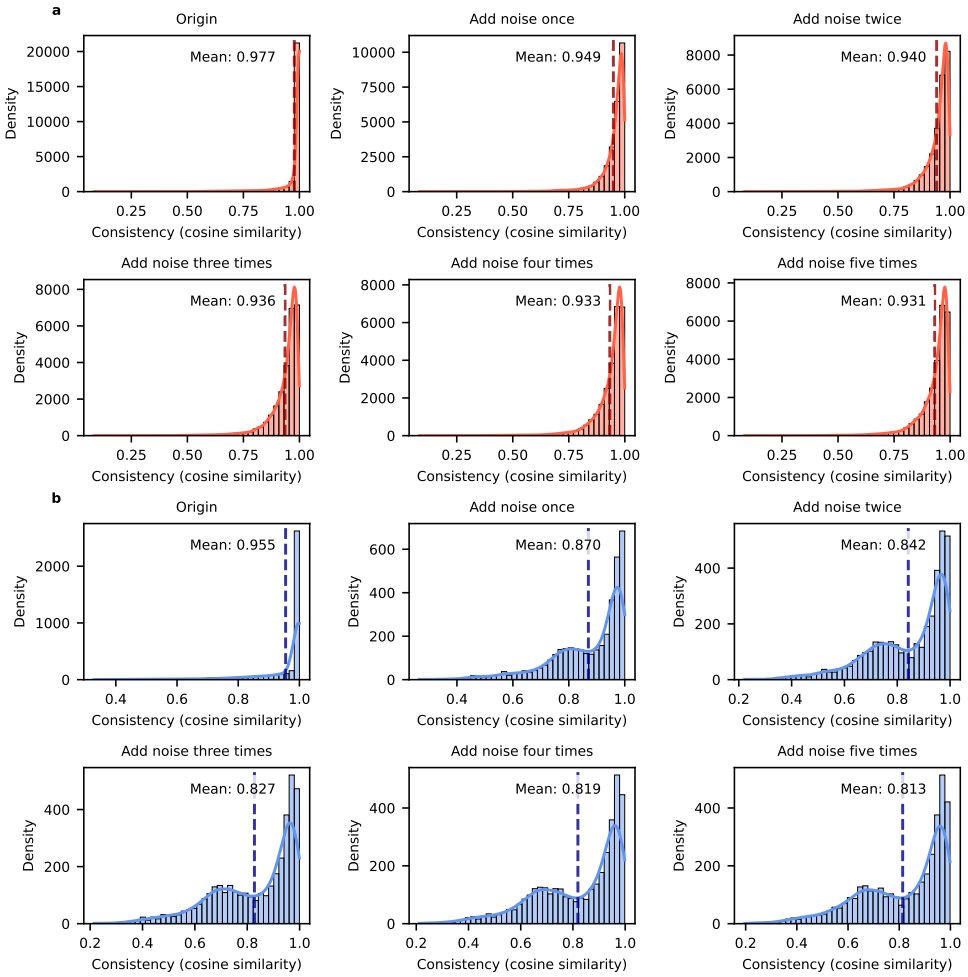
Extended Fig. 1: Performance on auxiliary tasks, including mass, atom number, and H/C ratio prediction. (a), (b), and (c) correspond to QTOF MS/MS, while (d), (e), and (f) correspond to Orbitrap MS/MS.



Extended Fig. 2: **Number of peaks and spectral entropy distributions.** The spectra under collision energy range $(20, 40]$ serve as an example. Spectra in other collision energy ranges reveal that $[\mathrm{M}+\mathrm{H}]^+$ and $[\mathrm{M}-\mathrm{H}]^-$ groups contain more complex spectra compared to $[2\,\mathrm{M}+\mathrm{H}]^+$ and $[2\,\mathrm{M}-\mathrm{H}]^-$ groups. The asterisk (***) denotes in the Mann-Whitney U test the p-value is smaller than 0.001, so the two hues are statistically significantly different.



Extended Fig. 3: **Potential impacts on performance on external test sets.** (a) Performance on compounds with varying similarity to the training set. For each compound in the test set, the maximum Tanimoto similarity with any compound in the training set is used as its compound similarity to the training set. FIDDLE's performance correlates with training set compound similarity, as its purely data-driven approach learns patterns without prior knowledge incorporation. (b) and (c) show the distributions of mass deviation and molecular mass in the test sets, respectively. It is worth noting that although the CASMI datasets were collected using QTOF instruments, we simulated the precursor m/z values as described in Section 4.1.2, resulting in a lower mass deviation compared to the EMBL dataset.



Extended Fig. 4: Consistency before and after data augmentation. (a) and (b) correspond to QTOF and Orbitrap, respectively. Note that data augmentation was not applied to Orbitrap in the experiments, as adding noise significantly decreases the MS/MS consistency for Orbitrap, and the training size for Orbitrap is quite large.