Efficiently predicting high resolution mass spectra with graph neural networks

Michael Murphy [*1,2] **Stefanie Jegelka** [1] **Ernest Fraenkel** [2] **Tobias Kind** [3] **David Healey** [3] **Thomas Butler** [3]

Abstract

Identifying a small molecule from its mass spectrum is the primary open problem in computational metabolomics. This is typically cast as information retrieval: an unknown spectrum is matched against spectra predicted computationally from a large database of chemical structures. However, current approaches to spectrum prediction model the output space in ways that force a tradeoff between capturing high resolution mass information and tractable learning. We resolve this tradeoff by casting spectrum prediction as a mapping from an input molecular graph to a probability distribution over chemical formulas. We further discover that a large corpus of mass spectra can be closely approximated using a fixed vocabulary constituting only 2% of all observed formulas. This enables efficient spectrum prediction using an architecture similar to graph classification – GRAFF-MS – achieving significantly lower prediction error and greater retrieval accuracy than previous approaches.

1. Introduction

The identification of unknown small molecules in complex mixtures is a primary challenge in many areas of chemical and biological science. The standard high-throughput approach to small molecule identification is *tandem mass spectrometry* (MS/MS), with diverse applications including metabolomics (Dettmer et al., 2006), drug discovery (Atanasov et al., 2021), clinical diagnostics (Evans et al., 2020), forensics (Brown et al., 2020), and environmental monitoring (Hernández et al., 2012).

MS/MS generates an experimental signature – a *mass spectrum* – of an unknown molecule by breaking it into fragments. The spectrum contains a (mass-to-charge ratio, height) tuple for each resulting fragment, reflecting its elemental composition, electric charge, and tendency to form. The problem of inferring the 2D structure of a molecule from its spectrum is known as *structural elucidation*. Structural elucidation is the primary computational bottleneck in MS/MS, and is far from solved: typically only $2{-}4\%$ of spectra are identified in untargeted metabolomics experiments (da Silva et al., 2015). A recent competition – the 2022 Critical Analysis of Small Molecule Identification (CASMI) challenge (Fiehn, 2022) – saw no more than 30% accuracy, with algorithmic approaches only marginally outperforming manual annotation by expert chemists.

Because MS/MS is a lossy measurement, and available training sets are small, direct prediction of structures from spectra is particularly challenging. The approach for small molecule identification preferred in practice by most users of mass spectrometry is *spectral library search*, which casts the problem as information retrieval (Stein, 2012): an observed spectrum is queried against a library of spectra with known structures. This provides an informative prior, and has the advantage of easy interpretability. But as there are relatively few (10^4) small molecules with publicly known experimental mass spectra, in spectral library search it is necessary to augment libraries with spectra predicted from large databases $(10^6 - 10^9)$ of molecular graphs. This motivates the problem of *spectrum prediction*: the ability to predict higher-quality spectra from large chemical structure databases could greatly increase compound identification rates in real experimental settings.

Spectrum prediction is actively studied in metabolomics and quantum chemistry (Krettler & Thallinger, 2021), yet has historically received little attention from the machine learning community. A major challenge in spectrum prediction is modelling the output space: a mass spectrum is a variable-length set of real-valued tuples, which is not straightforward to represent as an output of a machine learning model. The mass-to-charge (m/z) coordinate poses particular difficulty: it must be predicted with high precision, as a key strength of MS/MS is the ability to distinguish small fractional m/z differences (on the order of 10^{-6}) representative of different elemental composition.

*The lead author carried out this work as an intern at Enveda Biosciences. [1]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA [2]Department of Biological Engineering, MIT, Cambridge, MA, USA [3]Enveda Biosciences, Boulder, CO, USA. Correspondence to: Michael Murphy <murphy17@mit.edu>, Thomas Butler <tom.butler@envedabio.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Previous approaches to spectrum prediction force a tradeoff between capturing high resolution m/z information and tractability of the learning problem. *Mass-binning* methods (Wei et al., 2019; Zhu et al., 2020; Young et al., 2021) represent a spectrum as a fixed-length vector by discretizing the m/z axis at regular intervals, discarding fine-scale information in favor of tractable learning. *Bond-breaking* methods (Wang et al., 2021; Ruttkies et al., 2019) achieve perfect m/z resolution, but use expensive combinatorial enumeration of substructures.

Our work presents a novel formulation that exploits the many-to-one relationship between molecular graphs and chemical formulas. Specifically, we make the following contributions:

- We formulate spectrum prediction as a mapping from a molecular graph to a probability distribution over chemical formulas, allowing full resolution predictions without enumerating substructures;
- We discover most mass spectra can be effectively approximated with a small fixed vocabulary of chemical formulas, bypassing the tradeoff between m/z resolution and tractable learning; and
- We implement an efficient graph neural network architecture, GRAFF-MS, that outperforms state-of-the-art in both prediction error and runtime on canonical mass spectrometry datasets, and yields superior accuracy on a large-scale structure retrieval task.

2. Background

We denote vectors $\boldsymbol{x}$ in bold lowercase and matrices $\boldsymbol{X}$ in bold uppercase.

A *molecular graph* $G = (V, E, \boldsymbol{a}, \boldsymbol{b})$ is a minimal description of the 2D structure of a molecule: it comprises an undirected graph with nodes V representing atoms, and edges $E \subset V \times V$ representing bonds. Each node i is labelled with a chemical element $a_i \in \{\text{C, H, N, O, P, S} \dots\}$, and each edge (i, j) with a bond order $b_{ij} \in \{1, 1.5, 2, 3\}$.

A *chemical formula* f (e.g. $C_8H_{10}N_4O_2$) describes a multiset of atoms, which we encode as a nonnegative integer vector of atom counts in $\mathcal{F}^* \doteq \mathbb{Z}_+^{\{\text{C, H, N, O, P, S} \dots\}}$. Formulas may be added and subtracted from one another, and inequalities between formulas are taken to hold elementwise. The *subformulas* of f are the set $\mathcal{F}(f) \doteq \{f' \in \mathcal{F}^* : f' \leq f\}$.

$\langle \mu, f \rangle \in \mathbb{R}_+$ is the *theoretical mass* of formula f, in units of daltons (Da): this is a weighted sum of the monoisotopic masses of the elements of the periodic table, $\mu \in \mathbb{R}_+^{\{\text{C, H, N, O, P, S} \dots\}}$, with multiplicities given by f.

A *mass spectrum* S is a variable-length set of *peaks*, each of which is a (m/z, intensity) tuple $(m_i, y_i) \in \mathbb{R}_+^2$. We use the notation $i \in S$ to index peaks in a spectrum. We assume spectra are normalized, permitting us to treat them as probability distributions: $\sum_{i \in S} y_i = 1$. A mass spectrum is implicitly always accompanied by a precursor formula P. We always assume charge $z = 1$, as it is rare for small molecules to acquire more than a single charge.

2.1. Tandem mass spectrometry

A tandem mass spectrometer is a scientific instrument that generates high-throughput experimental signatures of the molecules present in a complex mixture. It operates by ionizing a chemical sample into a jet of electrically-charged gas. This gas is electromagnetically filtered to select a population of *precursor ions* of a specific mass-to-charge ratio (m/z) representing a unique molecular structure. Each precursor ion is fragmented by collision with molecules of an inert gas. If a collision occurs with sufficient energy, one or more bonds in the precursor will break, yielding a charged *product ion* and one or more uncharged *neutral loss* molecules. The product ion is measured by a detector, which records its m/z up to a small error proportional to m/z times the instrument resolution ϵ, typically on the order of 10^{-6}. This process is repeated for large numbers of identical precursor ions, building up a histogram indexed by m/z. Local maxima in this histogram are termed *peaks*: ideally, each peak represents a unique product ion, with intensity reflecting its probability of formation. This set of peaks constitutes the mass spectrum. A typical mass spectrometry experiment acquires mass spectra for tens of thousands of distinct precursors in this manner. We depict this process in Figure 1A, and the relationship between precursor ion, product ion and neutral loss in Figure 1B.

2.2. Mass decomposition

Modern mass spectrometry achieves sufficiently high resolution to detect small deviations in m/z from integrality that are characteristic of different chemical elements. This property is a key strength of the technology, because it permits annotating peaks with formulas through *mass decomposition* (Dührkop et al., 2013). Given a product ion of $m/z = m$, a precursor formula P, and an instrument resolution ϵ, *product mass decomposition* yields a set $\mathcal{F}(P, m, \epsilon)$ of chemically plausible subformulas of P whose theoretical masses lie within the (multiplicative) measurement error ϵm of m. This can be cast as the following integer program, in which all solutions with cost $\leq \epsilon m$ are enumerated:

$$\min_{f \in \mathcal{F}^*} |\langle \mu, f \rangle - m| \tag{1}$$

$$\text{s.t. } f \leq P, f \in \Omega \tag{2}$$

where Ω describes a general set of constraints that exclude unrealistic chemical formulas (Kind & Fiehn, 2007). For

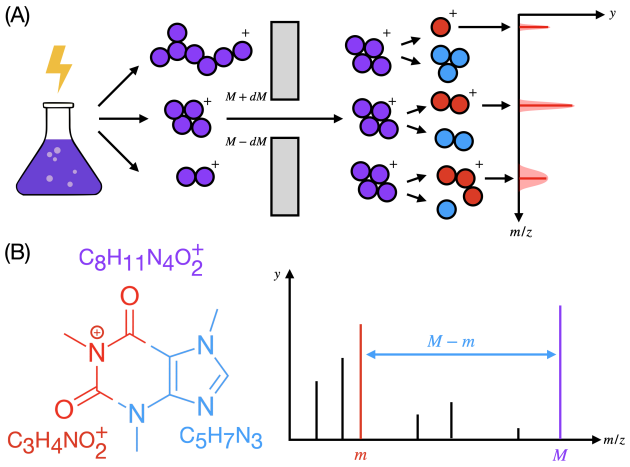

Figure 1. (A) The workflow of tandem mass spectrometry. A chemical mixture is ionized and filtered to isolate precursor ions of $m/z = M$; these are fragmented into product ions (red) and neutral losses (blue), and a detector yields a histogram of product ions indexed by m/z, with measurement error proportional to m/z. (B) An example fragmentation of a precursor ion with formula $C_8H_{11}N_4O_2$. Fragmentation breaks bonds, cutting the molecular graph into connected components. The component retaining the charge is the product ion; its complement is the neutral loss. The peak at $m/z = m$ represents the product ion; given the precursor formula, we can equally specify this peak by its formula $C_3H_4NO_2$, or the formula of its neutral loss $C_5H_7N_3$.

modern instruments, ϵ is sufficiently small for there to typically be only one or a few valid solutions. This allows us to later rely on product mass decomposition as a black-box to generate useful formula annotations at training time.

3. Related work

Bond-breaking is the most studied approach to spectrum prediction (Allen et al., 2014; Wang et al., 2021; Ruttkies et al., 2019; Cao et al., 2021). This solves the problem of representing the output space by enumerating the structures of all probable product ions: these are taken to be connected subgraphs of the precursor, generated by sequences of edge removals. Each product ion structure is scored for its probability of formation, and a spectrum is generated by associating this probability with each structure's theoretical m/z. Bond-breaking therefore achieves perfect m/z resolution, but suffers from two major weaknesses: first, enumerating substructures scales poorly with molecule size, and is not conducive to massively-parallel implementation on a GPU. We found a state-of-the-art method (Wang et al., 2021) takes ∼5s on average to predict a single mass spectrum, which precludes training on the largest available datasets: using the same settings as its authors, training (Wang et al., 2021) on ∼300k spectra in NIST-20 would take an estimated *three months* on a 64-core machine. It also poses serious limitations at test time, as inference with a large-scale structure database like ChEMBL (Gaulton et al., 2016) requires predicting millions of spectra. The other weakness of bond-breaking arises from a restrictive modelling assumption: *rearrangement reactions* (McLafferty, 1959) frequently yield product ions that are not reachable from the precursor by sequences of edge removals.[1]

Mass-binning is used for spectrum prediction by (Wei et al., 2019), and subsequently employed in recent preprints (Zhu et al., 2020; Young et al., 2021). This approach represents a mass spectrum as a fixed-length vector via discretization: the *m/z* axis is partitioned into narrow regularly-spaced bins, and each bin is assigned the sum of the intensities of all peaks falling within its endpoints. Spectrum prediction then becomes a vector-valued regression problem, which is conducive to GPU implementation and scales better than bond-breaking. But because a target space with millions of mass bins is too large, realistic bin counts lose essential high resolution information about the chemical formulas of the peaks: discarding a key strength of MS/MS analysis in favor of a tractable learning problem. Such models are also susceptible to edge effects, where m/z measurement error of the instrument can cause peaks of the same product ion to cross bin boundaries from spectrum to spectrum.

Other approaches include molecular dynamics simulation (Koopman & Grimme, 2021), which has extremely high computational costs; and NLP-inspired models for peptides (Zhou et al., 2017; Gessulat et al., 2019), which are effective but inapplicable to other types of molecules.

While this manuscript was under review, two GNN-based approaches to modelling mass spectra as distributions over subformulas were also published. Rather than the fixed vocabulary approximation we discover here, Zhu & Jonas (2023) use an exhaustive enumeration scheme to generate subformulas, which are then used in a formula-to-atom attention operation to predict a peak height for each. Goldman et al. (2023) decode a prefix-tree of plausible subformulas from the molecular graph and predict intensities for these using a set-to-set transformer.

4. GRAFF-MS

Our approach comprises three major components: first, we represent the output space of spectrum prediction as a space of probability distributions over chemical formulas. We then introduce a constant-sized approximation of this output space using a fixed vocabulary of formulas, which we can generate from our training data; we later show this in-

[1]While engineered rules are used in bond-breaking to account for certain well-studied rearrangements, we found the state-of-the-art method CFM-ID still fails to assign a formula annotation within 10ppm to 42% of monoisotopic peaks in the NIST-20 dataset.

troduces only minor approximation cost, as most formulas occur with low probability. Finally, we derive a loss function that takes into account data-specific ambiguities introduced by our model of the output space. These components together allow us to efficiently predict spectra using a standard graph neural network architecture.

We call our approach GRAFF-MS: (Gr)aph neural network for (A)pproximation via (F)ixed (F)ormulas of (M)ass (S)pectra.

4.1. Modelling spectra as probability distributions over chemical subformulas of the precursor

Our aim is to predict a mass spectrum from a molecular graph. To do so, we must determine how to best represent the output space: a spectrum consists of a variable-length set of peaks located at continuous m/z positions, whose heights sum to one. We notice that peaks are not located arbitrarily: the set of m/zs is structured, as the m/z of a peak is determined (up to measurement error) by the chemical formula of its corresponding product ion. This formula is sufficient to determine the m/z; in particular, we do not need to know the product ion's full 2D structure. We therefore model a mass spectrum as a probability distribution over *chemical subformulas* $\mathcal{F}(P)$ of the precursor P:

$$S = \{(m_f, y_f) : f \in \mathcal{F}(P)\} \tag{3}$$

where $m_f \doteq \langle \mu, f \rangle$ is the theoretical mass of formula f.

This is more efficient in principle than bond-breaking, which models a spectrum as a distribution over at worst exponentially many *substructures* of the precursor. In contrast, the number of *subformulas* is only polynomial in the coefficients of the precursor formula – and the majority of subformulas can be ruled out *a priori* as chemically infeasible (Kind & Fiehn, 2007). It is also less restrictive than bond-breaking, which relies on hand-engineered rules to capture rearrangement reactions: enumerating subformulas is guaranteed to cover all possible peaks, irrespective of whether the structure of their product ion is reachable by edge removals or not. Yet our approach preserves the core advantage of bond-breaking over mass-binning: predicting a height for each subformula yields spectra with perfect m/z resolution.

4.2. Fixed vocabulary approximation of formula space

In practice, enumerating subformulas is still a costly operation for larger molecules. One way to avoid this would be to sequentially decode formulas of nonzero probability one at a time: we opt not to do so, as this requires a more complex, data-hungrier model, and necessitates a linear ordering of formulas, for which there is not an obvious correct choice. Instead, we exploit a property of small molecule mass spectra that we discovered in this work and illustrate in Figure 2: almost all of the signal in small molecule mass spectra lies in peaks that can be explained by a relatively small number ($\sim$2%) of product ion and neutral loss formulas that frequently recur across spectra.

Inspired by this finding, we approximate $\mathcal{F}(P)$ via the union $\hat{\mathcal{F}}(P) = \hat{\mathcal{P}} \cup (P - \hat{\mathcal{L}})$ of a fixed set of frequent product ion formulas $\hat{\mathcal{P}}$ and a variable set of 'precursor-specific' formulas $P - \hat{\mathcal{L}}$ obtained by subtracting a fixed set of frequent neutral loss formulas $\hat{\mathcal{L}}$ from the precursor P. This greatly simplifies the spectrum prediction problem: we now only need to predict a probability for each of the formulas in $\hat{\mathcal{P}}$ and $\hat{\mathcal{L}}$, which we can accomplish with time *constant* in the size of the precursor.

Stated explicitly, we approximate the spectrum as:

$$S \approx \{(m_f, y_f) : f \in \hat{\mathcal{F}}(P)\} \tag{4}$$

where a height of zero is implicitly assigned to any formula not in $\hat{\mathcal{F}}(P)$.

The fact that we can equally represent a product ion by either its own formula or a neutral loss formula relative to its precursor ion is crucial to generalization, as also noted by Wei et al. (2019). If we only included frequent product ion formulas, we would explain peaks of low mass well, which typically correspond to small charged functional groups. But as formula space becomes larger with increasing mass, it becomes increasingly unlikely that every significant peak of higher mass in an unseen compound will be explained. However, such peaks do not represent arbitrary subformulas of the precursor: they tend to arise from losses of small uncharged functional groups and combinations thereof, which we capture by including frequent neutral losses.

Our algorithm to generate $\hat{\mathcal{P}}$ and $\hat{\mathcal{L}}$ involves listing all product ion and neutral loss formulas yielded by mass decomposition of the training set, and ranking them by the sum of the heights of all peaks to which each formula is assigned; we select the top K highest ranked among either type. The algorithm is provided in appendix A.

4.3. Peak-marginal cross entropy

To train our approach, we must rely on formula annotations generated by mass decomposition. Because mass spectrometers have limited resolution, often more than one valid subformula has a mass within measurement error of a peak. These are considered equiprobable *a priori*, and need not be mutually exclusive: it is possible for a compound to contain two distinct substructures with m/z difference smaller than the measurement error. As we cannot pick a single formula in such cases, we approximate the full cross entropy loss by marginalizing over compatible formulas: we term this the *peak-marginal cross entropy*. We minimize this loss with

respect to the parameters of a neural network $\hat{\boldsymbol{y}}(\cdot;\theta)$:

$$\min_{\theta} -\sum_{n=1}^{N}\sum_{i\in S_n} y_i^n \log \sum_{f\in\hat{\mathcal{F}}_i^n} \hat{y}_f(G_n;\theta) \tag{5}$$

using $\hat{\mathcal{F}}_i^n \doteq \hat{\mathcal{F}}(P_n)\cap\mathcal{F}(P_n, m_i^n, \epsilon)$ to indicate the intersection of our fixed vocabulary with the formula annotations for peak i of spectrum n. We provide a derivation from first principles in appendix B.

In this formulation, given a molecular graph G of a precursor with formula P, our model predicts a probability $\hat{y}_f$ for every formula f in the fixed vocabulary. This produces a spectrum $\hat{S} = \{(m_f, \hat{y}_f) : f \in \hat{\mathcal{F}}(P)\}$. These per-formula probabilities are summed within each observed peak across its compatible formulas to yield a predicted peak height, and the cross-entropy between the observed and predicted peak heights across the entire spectrum is minimized.

4.4. Model architecture

Formulating spectrum prediction as graph classification permits applying a typical GNN architecture. GRAFF-MS uses a graph isomorphism network with edge and graph Laplacian features (Xu et al., 2019; Hu et al., 2020; Lim et al., 2022). This encodes the molecular graph by a dense vector representation, which is then conditioned on mass-spectral covariates and passed through a feed-forward network that decodes a logit for each formula in the vocabulary.

We start with the graph of the 2D structure $G = (V, E)$, to which we add a virtual node (Gilmer et al., 2017) and four classes of features: node features $\boldsymbol{a}_i \in \mathbb{R}^{d_{atom}}$, edge features $\boldsymbol{b}_{ij} \in \mathbb{R}^{d_{bond}}$, covariate features $\boldsymbol{c} \in \mathbb{R}^{d_{cov}}$, and the top eigenvectors and eigenvalues of the graph Laplacian $\boldsymbol{v}_i \in \mathbb{R}^{d_{eig}}, \boldsymbol{\lambda} \in \mathbb{R}^{d_{eig}}$. We use the canonical atom and bond featurizers from `DGL-LifeSci` (Li et al., 2021) to generate $\boldsymbol{a}$ and $\boldsymbol{b}$. Since a mass spectrum is not fully determined by the molecular graph, $\boldsymbol{c}$ includes a number of necessary experimental parameters: normalized collision energy, precursor ion type, instrument model, and presence of isotopic peaks. Further details are provided in Table 3 in the appendix; there we also provide hyperparameter settings.

We first embed the node, edge, and covariate features into $\mathbb{R}^{d_{enc}}$, reusing the following `MLP` block:

$$\texttt{MLP}(\cdot) = \texttt{LayerNorm}(\texttt{Dropout}(\texttt{SiLU}(\texttt{Linear}(\cdot))))$$

and transform the Laplacian features into node positional encodings in $\mathbb{R}^{d_{enc}}$ using a SignNet (Lim et al., 2022) with ϕ and ρ both implemented as 2-layer stacked `MLP`s:

$$\boldsymbol{x}_i^{atom} = \texttt{MLP}_{atom}(\boldsymbol{a}_i) \tag{6}$$

$$\boldsymbol{x}_i^{eig} = \texttt{SignNet}(\boldsymbol{v}_i, \boldsymbol{\lambda}) \tag{7}$$

$$\boldsymbol{x}_{ij}^{bond} = \texttt{MLP}_{bond}(\boldsymbol{b}_{ij}) \tag{8}$$

$$\boldsymbol{x}^{cov} = \texttt{MLP}_{cov}(\boldsymbol{c}) \tag{9}$$

taking $i \in V$ and $(i,j) \in E$. We sum the embedded atom features and node positional encodings, and pass these along with embedded bond features into a stack of alternating L message-passing layers to update the node representations, and L `MLP` layers to update the edge representations.

$$\boldsymbol{x}_i^{(0)} = \boldsymbol{x}_i^{atom} + \boldsymbol{x}_i^{eig} \tag{10}$$

$$\boldsymbol{e}_{ij}^{(0)} = \boldsymbol{x}_{ij}^{bond} \tag{11}$$

$$\boldsymbol{X}^{(l+1)} = \boldsymbol{X}^{(l)} + \texttt{GINEConv}^{(l)}(G, \boldsymbol{X}^{(l)}, \boldsymbol{E}^{(l)}) \tag{12}$$

$$\boldsymbol{e}_{ij}^{(l+1)} = \boldsymbol{e}_{ij}^{(l)} + \texttt{MLP}_{edge}^{(l)}(\boldsymbol{e}_{ij}^{(l)} \| \boldsymbol{x}_i^{(l+1)} \| \boldsymbol{x}_j^{(l+1)}) \tag{13}$$

where $\|$ denotes concatenation. The message-passing layer uses the `GINEConv` operation implemented in (Fey & Lenssen, 2019): for its internal feed-forward network, we use two stacked `MLP` blocks with GraphNorm (Cai et al., 2021) in place of layer normalization. We similarly replace layer normalization with GraphNorm in the $\texttt{MLP}_{edge}$ blocks. Both node and edge updates use residual connections, which we found greatly accelerate training.

We generate a dense representation of the molecule by attention pooling over nodes (Er et al., 2016), to which we add the embedded covariate features. An MLP decodes this into a spectrum representation $\boldsymbol{x}^{spec} \in \mathbb{R}^{d_{dec}}$:

$$a_i = \texttt{Softmax}_{i\in V}(\texttt{Linear}(\boldsymbol{x}_i)) \tag{14}$$

$$\boldsymbol{x}^{mol} = \sum\nolimits_{i\in V} a_i \boldsymbol{x}_i^{(L)} \tag{15}$$

$$\boldsymbol{x}^{spec} = \texttt{MLP}_{spec}(\boldsymbol{x}^{mol} + \boldsymbol{x}^{cov}), \tag{16}$$

where $\texttt{MLP}_{spec}$ is a stack of L' `MLP` blocks with residual connections. In principle, we may now project this representation via a linear layer $(\boldsymbol{w}_k, b_k)$ into a logit z_k for each of the K product ion or neutral loss formulas in the vocabulary.

4.5. Domain-specific modifications

We must now introduce some corrections motivated by domain knowledge to produce realistic mass spectra.

Depending on instrument parameters, tandem mass spectra can display small peaks arising from higher isotopic states of the precursor ion, at integral m/z shifts relative to the monoisotopic peak. We model this as a source of noise: rather than expanding our vocabulary, we apply to all predictions a scalar offset for each isotopic state $\delta \in \{0, 1, 2\}$, which we parameterize as a linear function of $\boldsymbol{x}^{spec}$.

As our vocabulary includes both product ions and neutral losses, we also face occasional *double-counting*: depending on the precursor P, there are cases where the same subformula f will be predicted both as a product ion ($f \in \hat{\mathcal{P}}$)

and a neutral loss ($P - f \in \hat{\mathcal{L}}$). In such cases we subtract a $\log(2)$ correction factor from both logits: this way the innermost summation in Equation (5) takes the average of their contributions instead of their sum.

Applying these corrections and softmaxing yields the final heights of the predicted mass spectrum $\hat{\boldsymbol{y}}$:

$$z_{k\delta} = \boldsymbol{w}_k \boldsymbol{x}^{spec} + \boldsymbol{w}_\delta \boldsymbol{x}^{spec} + b_k - \mathbb{I}[k \text{ is double-counted}] \log(2) \quad (17)$$

$$\hat{y}_{k\delta} = \texttt{Softmax}_{k\delta}(z_{k\delta}) \quad (18)$$

where the subscript k is an index into our fixed vocabulary.

5. Experiments

5.1. Datasets

5.1.1. NIST-20

We train our model on the NIST-20 tandem MS library (National Institute of Standards and Technology, 2020). This is the largest commercial dataset of high resolution mass spectra of small molecules, curated by expert chemists, and is available for a modest fee.[2] For each measured compound, NIST-20 provides typically several spectra acquired across a range of collision energies. Each spectrum is represented as a list of (m/z, intensity, annotation) peak tuples, in addition to metadata describing instrumental parameters and compound identity. The annotation field includes a list of formula hypotheses per peak that were computed by NIST using mass decomposition and verified by expert chemists.

We restrict NIST-20 to HCD Orbitrap spectra with $[M + H]^+$ or $[M-H]^-$ precursor ions. We exclude structures that are annotated as glycans or peptides or exceed 1000Da in mass (as these are not typically considered small molecules) or have atoms other than {C, H, N, O, P, S, F, Cl, Br, I}.

We use an $80/10/10$ structure-disjoint train/validation/test split, which we generate by grouping spectra according to the connectivity substring of their InChIKey (Heller et al., 2015), and assigning groups of spectra to splits. As the baseline CFM-ID only predicts monoisotopic spectra at qualitative energy levels {low, medium, high}, we restrict the test set to spectra with corresponding energies $\{20, 35, 50\}$ in which no peaks were annotated as higher isotopes. This yields 287,995 (18,665) training, 36,265 (2,346) validation, and 4,424 (1,632) test spectra (structures).

[2]Open data is not the norm in small molecule mass spectrometry, as large-scale annotated data has commercial value and requires substantial time commitment from teams of highly-trained human experts. No public-domain dataset comparable to NIST-20 therefore exists. However, NIST-20 and its predecessors are commonplace in academic mass spectrometry, and have been used in ML research (Wei et al., 2019; Dührkop, 2022).

5.1.2. CASMI-16

It is well known that uniform train-test splitting can overestimate generalization in molecular machine learning (Wu et al., 2018). To address this issue, we employ an independent test set: the spectra of the 2016 CASMI challenge (Schymanski et al., 2017). This is a small public-domain mass spectrometry dataset, constructed by domain experts specifically for benchmarking algorithms, and comprises structures selected as representative of those encountered 'in the wild' when performing mass spectrometry of small molecules.

We use $[M+H]^+$ and $[M-H]^-$ spectra from the combined 'Training' and 'Challenge' splits from Categories 2 and 3 of the challenge. We exclude any structures from CASMI-16 with an InChIKey connectivity match to any in NIST-20, yielding 166 spectra of 151 structures. CASMI-16 spectra are acquired with collision energy stepping, which generates a mixed spectrum from energies of $\{20, 35, 50\}$; for all methods we approximate this by predicting only the middle energy.

5.1.3. GNPS

To simulate performance in a real experimental setting, we extracted a subset of spectra from GNPS (Wang et al., 2016) that represent natural product molecules not found in NIST-20. This is a challenging dataset, as GNPS spectra are contributed by the community in an uncurated manner, and often are missing key covariates for spectrum prediction. To exclude obvious poor-quality spectra, we only consider $[M + H]^+$ and $[M - H]^-$ Orbitrap spectra, with reported precursor m/z matching the theoretical mass. GNPS does not report collision energy; we assume energy $= 35$, and only include spectra with a (number of peaks) to (precursor m/z) ratio between the 10$^{\text{th}}$ and 90$^{\text{th}}$ percentiles of that quantity for NIST-20 spectra acquired at that energy. This results in 677 mass spectra of 606 structures.

5.2. Baselines

5.2.1. CFM-ID

CFM-ID (Wang et al., 2021) is a bond-breaking method, viewed by the mass spectrometry community as the state-of-the-art in spectrum prediction (Krettler & Thallinger, 2021). We found CFM-ID prohibitively expensive to train on NIST-20 (one parallelized EM iterate on a subset of ∼60k spectra took 10 hours on a 64-core machine) so we use trained weights provided by its authors, learned from 18,282 spectra in the commercial METLIN dataset (Guijas et al., 2018). Domain experts consider spectra acquired under METLIN's conditions interchangeable with those of NIST-20 (Leoz et al., 2018) so it is reasonable to evaluate their model on our data.

5.2.2. NEIMS

NEIMS (Wei et al., 2019) is a feed-forward network that inputs a precomputed molecular fingerprint and outputs a mass-binned spectrum, which is postprocessed using a domain-specific gating operation ("bidirectional prediction"). As NEIMS was originally developed for electron-impact mass spectra, we retrained NEIMS on NIST-20, which necessitated two modifications: (1) we concatenate a vector of covariates to the fingerprint vector, without which NIST-20 spectra are not fully determined; and (2) we bin at 0.1Da intervals instead of 1Da intervals, to account for finer instrument resolution in NIST-20. We otherwise use the same hyperparameter settings as the original paper, and early-stop on validation loss.

5.3. Evaluation

5.3.1. Cosine similarity

We evaluate predictive accuracy against ground truth spectra via *mass-spectral cosine similarity* (Stein & Scott, 1994). This modifies the standard cosine similarity to allow for inexact matches in the m/z coordinates between two spectra. For two spectra S and $\hat{S}$, mass-spectral cosine similarity $C_{S,\hat{S}}$ is the maximal cosine similarity between vectors of peak heights taken over all peak matchings within a mass tolerance window τ. It is computed by solving a linear sum assignment problem:

$$C_{S,\hat{S}} \doteq \max_{x_{ij}\in\{0,1\}} \sum_{\substack{i\in S,\ j\in\hat{S}:\\ |m_i-\hat{m}_j|\le\tau}} x_{ij} \frac{y_i}{\|y\|_2} \frac{\hat{y}_j}{\|\hat{y}\|_2} \tag{19}$$

$$\text{s.t.} \ \sum_{i\in S} x_{ij} \le 1 \tag{20}$$

$$\sum_{j\in\hat{S}} x_{ij} \le 1 \tag{21}$$

We use the `CosineHungarian` implementation from `matchms` (Huber et al., 2020), with tolerance $\tau = 0.1$Da.

5.3.2. Time complexity

We also empirically compare dependence of runtime on input size between GRAFF-MS and the bond-breaking method CFM-ID. For fair comparison, we time a forward pass for each structure in the NIST-20 test split using only the CPU, without any batching. We include time spent in preprocessing: our input is a SMILES string and experimental covariates, and our output is a spectrum. As collision energy affects the number of fragments that CFM-ID generates, we predict spectra at low, medium, and high energies and use the average runtime of the three.

5.3.3. Spectral library search

We characterize retrieval performance on a large-scale spectral library search task. For each method, we predict a library of mass spectra from the structures in NIST-20, which we augment with 200k decoy structures sampled from ChEMBL within ± 0.1Da of any NIST-20 structure. $[M+H]^+$ and $[M-H]^-$ spectra are predicted at collision energies $\{20, 35, 50\}$, resulting in a library of 1,262,025 spectra of 221,502 structures. We query each of the 4,424 experimental spectra from the NIST-20 test split against this library, restricting comparisons to spectra with the same ionization mode and collision energy, of theoretical m/z within 0.1Da of the query. We rank the resulting matches by mass-spectral cosine similarity, and compute recall-at-k: both of the correct 2D structure, and of any 2D structure with the correct chemical formula.

6. Results

6.1. A fixed vocabulary of products and losses captures the vast majority of fragmentation events

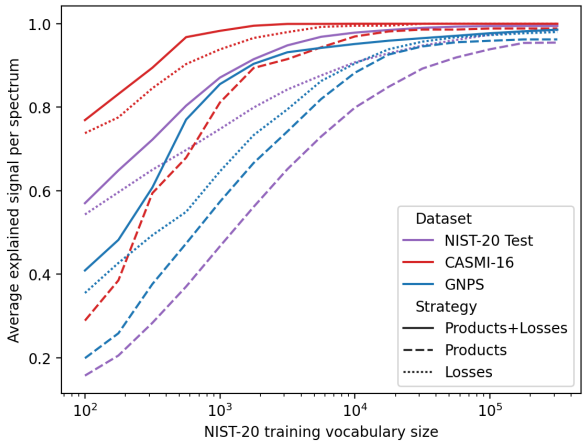

Figure 2. Generalization of different heuristics for fixed-size vocabulary selection. For a given vocabulary size on the x-axis, the y-axis indicates the sum of all explained peaks' heights within a given spectrum, averaged over all spectra.

Figure 2 illustrates the fraction of ion counts explained on average across mass spectra as the vocabulary size is varied. We observe most signal lies within peaks explainable by a relatively small number of product ion and neutral loss formulas. In particular, the vocabulary of $K = 10^4$ formulas we select from the NIST-20 training split (which contains 188,349 unique product ion and 351,165 unique neutral loss formulas) is sufficient to explain 98% of ion counts in the structure-disjoint test split. This vocabulary generalizes beyond NIST-20 to both CASMI-16 and GNPS, suggesting this 'formula sparsity' is a general property of small molecule mass spectra.

We also compare alternative strategies of picking only the top product ions or top neutral losses: our approach of using both types of formulas explains more signal for a fixed K than either type alone.

6.2. GRAFF-MS outperforms bond-breaking and mass-binning on standard MS/MS datasets

Table 1 shows GRAFF-MS produces spectra with greater cosine similarity to ground-truth than either baseline. These results hold for our test split of NIST-20 and for the independent test sets CASMI-16 and GNPS. We see all methods perform better on CASMI-16 than NIST-20: this is likely because NIST-20 includes a minority of substantially larger molecules (max weight 995Da) than CASMI-16 (max weight 539Da), with which all three methods struggle. Conversely, the poorer performance of all methods on GNPS likely reflects the difficulty of predicting the noisier spectra acquired in real-world biological experiments, as compared to curated spectra generated from libraries of pure chemical standards.

Table 1. Mean cosine similarity between predicted and true spectra on the NIST-20 test split, CASMI-16, and GNPS. 95% confidence intervals are computed via nonparametric bootstrap.

	NIST-20 TEST ($N = 4424$)	CASMI-16 ($N = 166$)	GNPS ($N = 677$)
CFM-ID	$0.53 \pm .01$	$0.71 \pm .04$	$0.26 \pm .02$
NEIMS	$0.60 \pm .01$	$0.57 \pm .06$	$0.29 \pm .02$
GRAFF-MS	$\mathbf{0.71} \pm .01$	$\mathbf{0.78} \pm .05$	$\mathbf{0.41} \pm .03$

To further characterize out-of-distribution performance, in Figure 3 we show how each method's performance on the NIST-20 test split decays as test examples decrease in similarity to the method's respective training set. Specifically, we compute, for each structure in the test set, the maximum Tanimoto similarity between its radius-2 Morgan fingerprint and that of any training structure. While all methods suffer out-of-distribution, GRAFF-MS maintains its edge over the other methods at all but the very highest levels of dissimilarity from NIST-20, where it approaches CFM-ID. CFM-ID's more gradual decay compared to the deep learning methods likely reflects the strong inductive bias of bond-breaking.

6.3. Representing peaks as subformulas scales better with molecular weight than substructures

Figure 4 shows how our approach to modelling high resolution spectra scales better with input size than bond-breaking. CFM-ID, which is written in optimized C++ code, takes on average 4.9 seconds per structure in the NIST-20 test split, and scales quadratically ($R^2 = 0.78$) with input size. (We believe this is because larger molecules in NIST-20 tend to be approximately path graphs – e.g. long hydrocarbon chains – with only quadratically many connected subgraphs.) In comparison, running our research implementation of GRAFF-MS on the CPU takes 1.3 core-seconds per spectrum, and scales approximately linearly ($R^2 = 0.65$). This pays off at larger molecular weight: for molecules

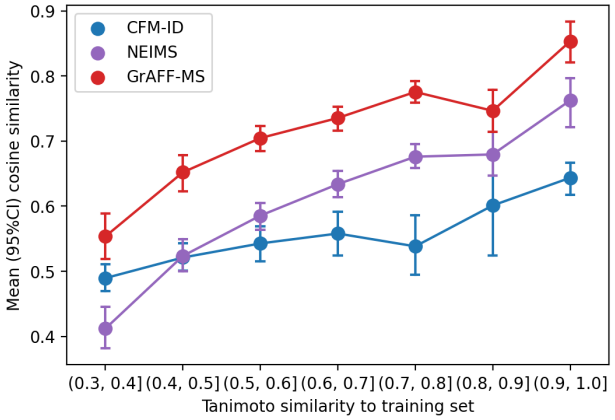

Figure 3. Relationship between predictive accuracy on NIST-20 test structures and structural similarity to the training set. Note that CFM-ID uses a different training set (METLIN).

$>$ 500Da, our model is 16$\times$ faster on average. Realistically, large-scale library prediction will use the GPU: on a single GPU with batch size 512, predicting all of the NIST-20 test spectra averages to 2.8ms per spectrum (mostly spent in CPU-bound preprocessing).

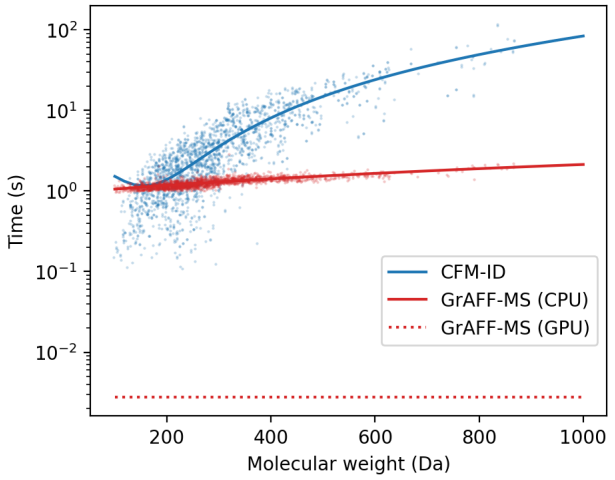

Figure 4. Empirical time complexity on NIST-20 structures with respect to molecular weight. Each dot is a structure. Solid lines are quadratic (blue) and linear (red) fits; dotted line indicates an average over all spectra computed using shuffled minibatches.

6.4. GRAFF-MS yields more accurate and faster structure retrieval against a large spectral library

In Table 2 we report recall at $k = \{1, 5, 10\}$ of experimental spectra from the NIST-20 test split on our spectral library search task. We see the superior performance of our approach in the spectrum prediction task extends to better downstream retrieval accuracy, in terms of both correct 2D structures and correct chemical formulas. Here the importance of speed at inference time becomes apparent: while

both deep learning methods took only a few minutes on a single GPU, CFM-ID required more than 5 node-days of compute time on 8 96-core nodes of an HPC cluster to generate this library.

Table 2. Recall-at-k of NIST-20 test spectra against synthetic libraries predicted from NIST-20 and ChEMBL structures. We report retrieval accuracy of both the correct 2D structure and of any structure with the correct chemical formula, as well as time taken to generate the 1.2 million spectra in the library.

	CFM-ID	NEIMS	GRAFF-MS
STRUCTURE, k=1	0.28±.01	0.27±.01	**0.37**±.01
5	0.56±.02	0.53±.02	**0.67**±.01
10	0.66±.01	0.61±.01	**0.75**±.01
FORMULA, k=1	0.34±.02	0.43±.01	**0.52**±.02
5	0.64±.02	0.65±.01	**0.76**±.01
10	0.74±.01	0.73±.01	**0.83**±.01
INFERENCE TIME	126H7M	**7M32S**	18M52S

6.5. GRAFF-MS distinguishes very similar compounds and makes human-like mistakes

In Figure 5 we show some particularly challenging examples of mass spectra. The top and middle panels show two structurally similar compounds, differing only by the order of one carbon-carbon bond. Our approach correctly predicts distinct spectra for each ($C_{S\hat{S}} = 0.90$, top; $C_{S\hat{S}} = 0.97$, middle). The third molecule is an example where we fail to predict a realistic spectrum ($C_{S\hat{S}} = 0.04$), but in a manner in which a human expert would also fail. This molecule is a member of the *phthalate* class, which chemists recognize by a characteristic dominant peak at 149Da (Jeilani et al., 2011). Our model predicts this same peak, correctly recognizing a phthalate. But in this case that peak is unusually minor – potentially reflecting a long-range dependency in the graph that our approach failed to capture.

7. Discussion

In this work, we develop GRAFF-MS, a graph neural network for predicting high resolution mass spectra of small molecules. Unlike previous approaches that force a trade-off between m/z resolution and a tractable learning problem, GRAFF-MS is both computationally efficient and capable of modelling the high resolution m/z information essential to modern mass spectrometry. This is made possible by our discovery that mass spectra of small molecules can be closely approximated as distributions over a fixed vocabulary of chemical formulas, highlighting the value that domain-aware modelling can add to molecular machine learning. Particularly surprising was that we outperform CFM-ID, which trades model expressivity for an even stronger scientific prior that we expected would contribute to better generalization. However, this prior incurs a heavy cost in time complexity, making it impractical to train CFM-ID on hundreds of thousands of spectra as we did.

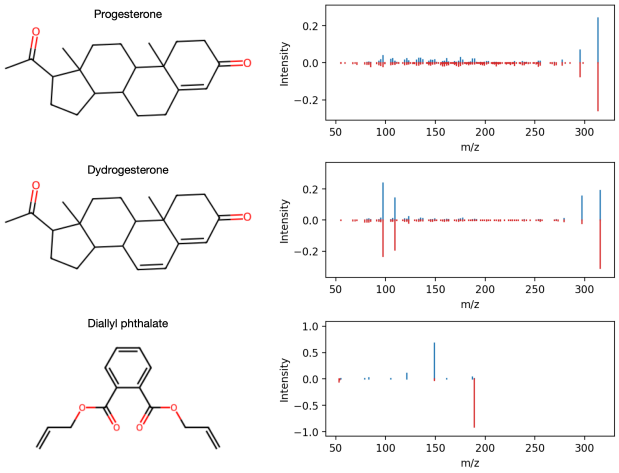

Figure 5. Three compounds from CASMI-16, with spectra predicted by our model (blue) against negated ground-truth (red). Oxygens are shaded red by convention.

While a fixed vocabulary of fragments yields an architecture that is simple to train and fast at inference time, it is possible that this can at times sacrifice flexibility: we may fail to capture fragments of intermediate size that can arise from complex small molecules such as natural products. We believe dynamic generation of this vocabulary, and the related problem of learning generalizable formula representations, to be promising avenues for future work.

Overall we anticipate GRAFF-MS will both accelerate scientific discovery and demonstrate mass spectrometry as a compelling domain for further machine learning research.

Software and Data

We provide code, data, and trained models at `https://github.com/murphy17/graff-ms`. The NIST-20 license agreement prohibits including spectra from it; we therefore provide instructions on how to obtain it.

Acknowledgements

The authors thank the reviewers for their constructive suggestions, and members of the Jegelka and Fraenkel labs, Gennady Voronov, Sam Goldman, and Connor Coley for helpful conversations. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported in this paper. M.M. thanks Enveda Biosciences, the Natural Sciences and Engineering Research Council of Canada, and the Chan-Zuckerberg Initiative for financial support.

References

Allen, F., Greiner, R., and Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, June 2014. doi: 10.1007/s11306-014-0676-4. URL https://doi.org/10.1007/s11306-014-0676-4.

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., and Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3): 200–216, 2021.

Brown, H. M., McDaniel, T. J., Fedick, P. W., and Mulligan, C. C. The current role of mass spectrometry in forensics and future prospects. *Analytical Methods*, 12(32):3974–3997, 2020. doi: 10.1039/d0ay01113d. URL https://doi.org/10.1039/d0ay01113d.

Cai, T., Luo, S., Xu, K., He, D., Liu, T., and Wang, L. Graphnorm: A principled approach to accelerating graph neural network training. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1204–1215. PMLR, 2021. URL http://proceedings.mlr.press/v139/cai21e.html.

Cao, L., Guler, M., Tagirdzhanov, A., Lee, Y.-Y., Gurevich, A., and Mohimani, H. Moldiscovery: learning mass spectrometry fragmentation of small molecules. *Nature Communications*, 12(1):1–13, 2021.

da Silva, R. R., Dorrestein, P. C., and Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112(41):12549–12550, 2015.

Dettmer, K., Aronov, P. A., and Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, August 2006. doi: 10.1002/mas.20108. URL https://doi.org/10.1002/mas.20108.

Dührkop, K. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra. *Bioinformatics*, 38(Supplement_1):i342–i349, June 2022. doi: 10.1093/bioinformatics/btac260. URL https://doi.org/10.1093/bioinformatics/btac260.

Dührkop, K., Ludwig, M., Meusel, M., and Böcker, S. Faster mass decomposition. In *Lecture Notes in Computer Science*, pp. 45–58. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40453-5_5. URL https://doi.org/10.1007/978-3-642-40453-5_5.

Er, M. J., Zhang, Y., Wang, N., and Pratama, M. Attention pooling-based convolutional neural network for sentence modelling. *Information Sciences*, 373:388–403, December 2016. doi: 10.1016/j.ins.2016.08.084. URL https://doi.org/10.1016/j.ins.2016.08.084.

Evans, E. D., Duvallet, C., Chu, N. D., Oberst, M. K., Murphy, M. A., Rockafellow, I., Sontag, D., and Alm, E. J. Predicting human health from biofluid-based metabolomics using machine learning. *Scientific Reports*, 10(1), October 2020. doi: 10.1038/s41598-020-74823-1. URL https://doi.org/10.1038/s41598-020-74823-1.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fiehn, O. Critical Assessment of Small Molecule Identification 2022. http://www.casmi-contest.org/2022/index.shtml/, 2022. [Online; accessed 12-December-2022].

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, November 2016. doi: 10.1093/nar/gkw1074. URL https://doi.org/10.1093/nar/gkw1074.

Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., and Wilhelm, M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509–518, May 2019. doi: 10.1038/s41592-019-0426-7. URL https://doi.org/10.1038/s41592-019-0426-7.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL http://proceedings.mlr.press/v70/gilmer17a.html.

Goldman, S., Li, J., and Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks, 2023. URL https://arxiv.org/abs/2304.13136.

Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A. E., Wolan, D. W., Spilker, M. E., Benton, H. P., and Siuzdak, G. METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry*, 90(5): 3156–3164, January 2018. doi: 10.1021/acs.analchem.7b04424. URL https://doi.org/10.1021/acs.analchem.7b04424.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7(1), May 2015. doi: 10.1186/s13321-015-0068-4. URL https://doi.org/10.1186/s13321-015-0068-4.

Hernández, F., Sancho, J. V., Ibáñez, M., Abad, E., Portolés, T., and Mattioli, L. Current use of high-resolution mass spectrometry in the environmental sciences. *Analytical and Bioanalytical Chemistry*, 403(5):1251–1264, February 2012. doi: 10.1007/s00216-012-5844-7. URL https://doi.org/10.1007/s00216-012-5844-7.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Huber, F., Verhoeven, S., Meijer, C., Spreeuw, H., Castilla, E., Geng, C., van der Hooft, J., Rogers, S., Belloum, A., Diblen, F., and Spaaks, J. matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, 5(52):2411, August 2020. doi: 10.21105/joss.02411. URL https://doi.org/10.21105/joss.02411.

Jeilani, Y. A., Cardelino, B. H., and Ibeanusi, V. M. Density functional theory and mass spectrometry of phthalate fragmentations mechanisms: Modeling hyperconjugated carbocation and radical cation complexes with neutral molecules. *Journal of the American Society for Mass Spectrometry*, 22(11), August 2011. doi: 10.1007/s13361-011-0215-8. URL https://doi.org/10.1007/s13361-011-0215-8.

Kind, T. and Fiehn, O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8(1), March 2007. doi: 10.1186/1471-2105-8-105. URL https://doi.org/10.1186/1471-2105-8-105.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. 2015.

Koopman, J. and Grimme, S. From QCEIMS to QCxMS: A tool to routinely calculate CID mass spectra using molecular dynamics. *Journal of the American Society for Mass Spectrometry*, 32(7):1735–1751, June 2021. doi: 10.1021/jasms.1c00098. URL https://doi.org/10.1021/jasms.1c00098.

Krettler, C. A. and Thallinger, G. G. A map of mass spectrometry-based iin silico/i fragmentation prediction and compound identification in metabolomics. *Briefings in Bioinformatics*, 22(6), March 2021. doi: 10.1093/bib/bbab073. URL https://doi.org/10.1093/bib/bbab073.

Leoz, M. L. A. D., Simón-Manso, Y., Woods, R. J., and Stein, S. E. Cross-ring fragmentation patterns in the tandem mass spectra of underivatized sialylated oligosaccharides and their special suitability for spectrum library searching. *Journal of the American Society for Mass Spectrometry*, 30(3):426–438, December 2018. doi: 10.1007/s13361-018-2106-8. URL https://doi.org/10.1007/s13361-018-2106-8.

Li, M., Zhou, J., Hu, J., Fan, W., Zhang, Y., Gu, Y., and Karypis, G. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega*, 2021.

Lim, D., Robinson, J., Zhao, L., Smidt, T., Sra, S., Maron, H., and Jegelka, S. Sign and basis invariant networks for spectral graph representation learning, 2022. URL https://arxiv.org/abs/2202.13013.

McLafferty, F. W. Mass spectrometric analysis. molecular rearrangements. *Analytical Chemistry*, 31(1):82–87, January 1959. doi: 10.1021/ac60145a015. URL https://doi.org/10.1021/ac60145a015.

National Institute of Standards and Technology. Nist-20, 2020. URL https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-\electron-ionization-spectral-libraries.

Ruttkies, C., Neumann, S., and Posch, S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics*, 20(1), July 2019. doi: 10.1186/s12859-019-2954-7. URL https://doi.org/10.1186/s12859-019-2954-7.

Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquière, B., and Neumann, S. Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9(1), March 2017. doi: 10.1186/s13321-017-0207-1. URL https://doi.org/10.1186/s13321-017-0207-1.

Stein, S. Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Analytical Chemistry*, 84(17):7274–7282, July 2012. doi: 10.1021/ac301205z. URL https://doi.org/10.1021/ac301205z.

Stein, S. E. and Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, September 1994. doi: 10.1016/1044-0305(94)87009-8. URL https://doi.org/10.1016/1044-0305(94)87009-8.

Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., and Wishart, D. S. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Analytical Chemistry*, 93(34):11692–11700, August 2021. doi: 10.1021/acs.analchem.1c01465. URL https://doi.org/10.1021/acs.analchem.1c01465.

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016.

Wei, J. N., Belanger, D., Adams, R. P., and Sculley, D. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Central Science*, 5(4):700–708, March 2019. doi: 10.1021/acscentsci.9b00085. URL https://doi.org/10.1021/acscentsci.9b00085.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/c7sc02664a. URL https://doi.org/10.1039/c7sc02664a.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Young, A., Wang, B., and Röst, H. Massformer: Tandem mass spectrum prediction with graph transformers, 2021. URL https://arxiv.org/abs/2111.04824.

Zhou, X.-X., Zeng, W.-F., Chi, H., Luo, C., Liu, C., Zhan, J., He, S.-M., and Zhang, Z. pDeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry*, 89(23):12690–12697, November 2017. doi: 10.1021/acs.analchem.7b02566. URL https://doi.org/10.1021/acs.analchem.7b02566.

Zhu, H., Liu, L., and Hassoun, S. Using graph neural networks for mass spectrometry prediction, 2020. URL https://arxiv.org/abs/2010.04661.

Zhu, R. L. and Jonas, E. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical Chemistry*, 95(5):2653–2663, January 2023. doi: 10.1021/acs.analchem.2c02093. URL https://doi.org/10.1021/acs.analchem.2c02093.

A. Fixed vocabulary selection

Algorithm 1 describes our procedure for selecting the product ions $\hat{\mathcal{P}}$ and neutral losses $\hat{\mathcal{L}}$. We use the shorthand $\mathcal{F}_i^n = \mathcal{F}(P_n, m_i, \epsilon)$ to indicate the set of formulas computed by product mass decomposition. When this yields more than one formula annotation for a peak, here we split the peak height uniformly among all annotations.

Algorithm 1 Fixed vocabulary selection

Input: training spectra and precursors $\{(S_n, P_n)\}_{n=1}^N$, vocabulary size K, tolerance ϵ
Output: product vocabulary $\hat{\mathcal{P}}$, loss vocabulary $\hat{\mathcal{L}}$
Initialize hash-tables $\Pi, \Lambda : \Pi(\cdot) = 0, \Lambda(\cdot) = 0$
Initialize sets $\hat{\mathcal{P}} = \emptyset, \hat{\mathcal{L}} = \emptyset$
for $n = 1 \dots N$ **do**
  for $(m_i, y_i) \in S_n$ **do**
    Compute mass decomposition $\mathcal{F}_i^n = \mathcal{F}(P_n, m_i, \epsilon)$
    for $f \in \mathcal{F}_i^n$ **do**
      $l = P_n - f$
      $\Pi(f) \leftarrow \Pi(f) + y_i/|\mathcal{F}_i^n|$
      $\Lambda(l) \leftarrow \Lambda(l) + y_i/|\mathcal{F}_i^n|$
    end for
  end for
end for
Sort Π and Λ in descending value order
while $|\hat{\mathcal{P}}| + |\hat{\mathcal{L}}| \leq K$ **do**
  $f =$ first element of Π
  $l =$ first element of Λ
  if $\Pi(f) > \Lambda(l)$ **then**
    Add f to $\hat{\mathcal{P}}$
    Remove f from Π
  else
    Add l to $\hat{\mathcal{L}}$
    Remove l from Λ
  end if
end while

(In our implementation, we do not actually compute the mass decomposition in the loop: we instead simply read off the annotations provided already by NIST.)

B. Derivation of peak-marginal cross entropy

We derive our loss function from physical first principles, making a number of minor modelling assumptions:

- The number of precursor ions accumulated in a spectrum is Poisson with rate λ.
- Each individual precursor ion is independently converted into fragment j with probability p_j.
- The instrument resolution parameter ϵ is sufficiently small that separate peaks do not overlap: there exists exactly one peak $i(j)$ for every $j : p_j > 0$ satisfying $|\langle \mu, f_j \rangle - m_i| \leq \epsilon m_i$ (where f_j is the chemical formula of fragment j).

By the splitting property, the number of ions of each fragment are independently Poisson with rate $\lambda_j = \lambda p_j$. By the merging property, the height of peak i is also a Poisson r.v. K_i with rate $\lambda_i = \sum_{j \in \mathcal{J}_i} \lambda_j$, where $\mathcal{J}_i$ denotes the set of fragments whose theoretical masses fall within the measurement error ϵm_i of peak i. The log-likelihood of peak height is (taking equality up to constants C w.r.t. p_j):

$$\log P(K_i = k_i) \tag{22}$$
$$= k_i \log \lambda_i - \lambda_i - \log k_i! \tag{23}$$
$$= k_i \log \left(\sum_{j \in \mathcal{J}_i} \lambda_j \right) - \left(\sum_{j \in \mathcal{J}_i} \lambda_j \right) + C \tag{24}$$
$$= k_i \log \left(\sum_{j \in \mathcal{J}_i} \lambda p_j \right) - \left(\sum_{j \in \mathcal{J}_i} \lambda p_j \right) \tag{25}$$
$$= k_i \log \lambda + k_i \log \left(\sum_{j \in \mathcal{J}_i} p_j \right) - \lambda \sum_{j \in \mathcal{J}_i} p_j \tag{26}$$
$$= C + k_i \log \left(\sum_{j \in \mathcal{J}_i} p_j \right) - \lambda \sum_{j \in \mathcal{J}_i} p_j \tag{27}$$

where (24) uses merging, and (25) uses splitting. Because each product ion is assigned to exactly one peak (no overlap), the peak heights $\{K_i : i \in S\}$ are independent. Let the total number of accumulated ions $K = \sum_{i \in S} k_i$ in spectrum S. Defining $y_i = k_i / K$:

$$\log P(\{K_i = k_i : i \in S\}) \tag{28}$$
$$= \sum_{i \in S} \log P(K_i = k_i) \tag{29}$$
$$= \sum_{i \in S} \left(k_i \log \left(\sum_{j \in \mathcal{J}_i} p_j \right) - \lambda \sum_{j \in \mathcal{J}_i} p_j \right) \tag{30}$$
$$= \sum_{i \in S} (K y_i) \log \left(\sum_{j \in \mathcal{J}_i} p_j \right) - \lambda \sum_{i \in S} \sum_{j \in \mathcal{J}_i} p_j \tag{31}$$
$$= K \sum_{i \in S} y_i \log \left(\sum_{j \in \mathcal{J}_i} p_j \right) - \lambda \cdot 1 \tag{32}$$
$$= C \sum_{i \in S} y_i \log \left(\sum_{j \in \mathcal{J}_i} p_j \right) + C' \tag{33}$$

where (32) again uses our assumption that every fragment is assigned to exactly one peak. Dropping the constants and negating the final term yields the peak-marginal cross-entropy loss for a single spectrum.

C. Model hyperparameters

We use a vocabulary of $K = 10{,}000$ formulas. We train an $L = 6$-layer encoder and $L' = 2$-layer decoder with $d_{enc} = 512$ and $d_{dec} = 1024$, resulting in 24.1 million trainable parameters. We use the $d_{eig} = 8$ lowest-frequency eigenvalues, truncating or padding with zeros. Dropout is applied at rate 0.1. We use a batch size of 512 and the Adam optimizer (Kingma & Ba, 2015) with learning rate 5×10^{-4} and weight decay 10^{-5}. We train for 100 epochs and use the model from the epoch with the lowest validation loss. All models are trained using PyTorch Lightning with automatic mixed precision on 2 Tesla V100 GPUs.

D. Mass spectral covariates

Table 3. Mass spectral covariates used in our model.

Feature	Range	Comment
Collision energy	$[0, 200]$	Thermo Scientific PSB104, "Normalized Collision Energy Technology"
Precursor type	$[M+H]^+$, $[M-H]^-$	Includes ionization mode & adduct composition
Instrument model	Orbitrap Fusion Lumos, Thermo Finnigan Elite Orbitrap, Thermo Finnigan Velos Orbitrap	Different limits of detection
Has isotopic peaks	False, True	Proxy for width setting of precursor mass filter