MULTI-SCALE SINUSOIDAL EMBEDDINGS ENABLE LEARNING ON HIGH RESOLUTION MASS SPECTROMETRY DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Small molecules in biological samples are studied to provide information about disease states, environmental toxins, natural product drug discovery, and many other applications. The primary window into the composition of small molecule mixtures is tandem mass spectrometry (MS2), which produces data that are of high sensitivity and part per million resolution. We adopt multi-scale sinusoidal embeddings of the mass data in MS2 designed to meet the challenge of learning from the full resolution of MS2 data. Using these embeddings, we provide a new state of the art model for spectral library search, the standard task for initial evaluation of MS2 data. We also investigate the task of chemical property prediction from MS2 data, that has natural applications in high-throughput MS2 experiments and show that an average R^2 of 80% for novel compounds can be achieved across 10 chemical properties prioritized by medicinal chemists. We use dimensionality reduction techniques and experiments with different floating point resolutions to show the essential role multi-scale sinusoidal embeddings play in learning from MS2 data.

1 INTRODUCTION

Tandem Mass Spectrometry (MS/MS) is one of the primary tools used in metabolomics, the characterization of the small molecule ($\lesssim 1{,}000\,\mathrm{Daltons}$) contents of complex biological samples. Tandem Mass Spectrometry works by measuring with very high precision the masses of molecules and their constituent fragments. While MS techniques are highly sensitive and extremely precise, inferring the identity of the molecules and their properties from the resulting mass spectra is one of metabolomics' primary bottlenecks (Dunn et al., 2013). Unlocking the information contained in mass spectra will lead to an improvement in a number of applications including characterization of disease pathways, development of new agrochemicals, improved forensics analysis, and the discovery of new drugs (Zhang et al., 2020).

Profiling unknown molecules with mass spectrometry consists of several steps. First, molecules of interest are ionized and separated by their mass to charge ratio (m/z), resulting in the MS1 spectrum. Then, individual "precursor" ions are fragmented, and the m/z's of the fragments are recorded in the same manner. The resulting spectrum contains the m/z's and intensities (together, the "peaks") of all resulting fragments, and is called the MS2 spectrum. See Glish & Vachet (2003).

In recent years, several machine learning methods have been developed to identify the structures and properties of small molecules from their mass spectra. These approaches (Huber et al., 2021a;b; Kutuzova et al., 2021; Litsa et al., 2021; Shrivastava et al., 2021; van der Hooft et al., 2016) historically discretize m/z (via tokenization or binning). However, the m/z values obtained in modern mass spectrometry experiments are collected with parts per million levels of precision. There are three critical reasons to expect that modeling m/z values as numeric quantities is the appropriate technique. First, we know that relevant chemical information is present at the millidalton level (Jones et al., 2004; Pourshahian & Limbach, 2008) and discretization schemes typically strip this information away. Second, mass differences between peaks represent fragmentation patterns, and are therefore relevant to understanding the molecule. Learning to utilize this information from tokens is exceedingly difficult. Finally, any laboratory measurement always comes with some degree of experimental error,

and discretization techniques are particularly susceptible to measurement errors. In light of this, we proceed by modeling m/z as a numerical value using sinusoidal embeddings, which we hypothesize will enable us to capture information across many orders of magnitude.

In this work, we apply a numerical representation of m/z that uses sinusoidal embeddings across multiple scales to retain the information content of MS2 data across its entire mass resolution. We use these embeddings to achieve state of the art accuracy on spectral library search (Stein & Scott, 1994), a task which identifies compounds by comparing mass spectra from experiments to spectra from previously characterized compounds. This task is the primary method used in standard metabolomics analyses, but is challenging because spectra from a given compound vary widely according to experimental conditions. We also investigate the task of predicting chemical properties relevant to drug discovery from MS2 data and apply the same modeling approach to this task. We achieve 80% average R^2 for out of sample molecules across 10 properties on this task, which is high enough to enable first-pass filtering and selection of candidate drug molecules in high-throughput experiments based solely on spectral data. Finally, we qualitatively demonstrate that the multi-scale sinusoidal m/z embeddings capture high resolution structure by studying UMAP (McInnes et al., 2018) projections of the embeddings.

1.1 Related work

Modeling numerical data in terms of sinusoidal functions has a long history in many scientific fields. In machine learning, sinusoidal embeddings are most commonly used to encode the discrete positions of natural language token inputs in transformer models, which are otherwise position-agnostic (Vaswani et al., 2017). Other work has used sinusoidal embeddings for multi-dimensional positional encoding in image recognition (Li et al., 2021a). In mass spectrometry, sinusoidal embeddings have also been described in the context of modeling protein sequences from mass values, either with (Qiao et al., 2019) or without (Yilmaz et al., 2022) an initial mass binning step.

For modeling m/z values in the field of metabolomics, previous machine learning models have relied primarily on discretization of the continuous mass inputs. This is usually accomplished by binning m/z values into fixed length vectors with peak intensity as the value for each element. Various authors have used binned representations of spectra for spectral library search (Huber et al., 2021b), unsupervised topic modeling (van der Hooft et al., 2016), and molecule identification (Kutuzova et al., 2021; Litsa et al., 2021). Alternatively, masses have been tokenized via rounding for tasks such as unsupervised spectral similarity (Huber et al., 2021a) and molecule prediction from synthetic data (Shrivastava et al., 2021).

Other discretization approaches rely on tokenization of m/z values by direct molecular formula assignment. Even when this method is able to uniquely identify molecular formulas for every fragment ion, it still faces a problem endogenous to tokenization: it's very difficult to reason about the relationships between discrete tokens except by pattern recognition over a very large number of training examples. Böcker & Rasche (2008) addressed this difficulty by modeling the fragmentation process with a tree structure, and Dührkop et al. (2015) used these fragmentation trees as inputs to a molecular fingerprint prediction model that is used to assign molecules to spectra from molecular structure databases.

2 Model

2.1 Base model architecture

An MS2 spectrum,

$$S = \left\{ (m/z, I)_{\text{precursor}}, (m/z, I)_{\text{fragment}_1}, \ldots, (m/z, I)_{\text{fragment}_N} \right\}, \tag{1}$$

is composed of a precursor m/z and a set of N fragment peaks at various m/z's and intensities (I). Various other data are typically collected, including precursor abundance, charge, collision energy, *etc*, but these are not used in this work. Transformer encoders (Vaswani et al., 2017) without the positional encoding are explicitly fully-symmetric functions and hence are ideally suited to model a set of fragmentation peaks. We therefore take our base model, $\text{SpectrumEncoder}(S)$, to be a transformer encoder, whose inputs are a set of $(m/z, I)$ pairs that includes the precursor along with

all of the fragment peaks. We normalize the intensities to a maximum of 1 for fragments and assign an intensity of 2 to the precursor. Finally, we take as output the embedding vector from the final transformer layer corresponding to the precursor input. See Figure 2. Note, the use of transformers to study synthetic MS2 spectra has been explored in Shrivastava et al. (2021).

This flexible approach enables us to experiment with and compare various representations of MS2 data. We pursue two approaches: tokenization of m/z and modeling m/z as numerical values via sinusoidal embeddings. Before describing these, we need the following definition for a simple two layer feed forward MLP,

$$\mathrm{FF}\left(x\right) = W_2 \mathrm{ReLu}\left(W_1 x + b_1\right) + b_2. \tag{2}$$

Note that each occurrence of FF we employ below is a separate instance with distinct weights.

2.1.1 TOKENIZED m/z PEAK EMBEDDING

Our first approach is to discretize m/z by rounding to 0.1 precision and treating these objects as tokens (as in an NLP context). These tokens are embedded in a dense vector space as in Huber et al. (2021a); Mikolov et al. (2013), via an embedding function TE. We then construct the token peak embedding,

$$\mathrm{PE}_{\mathrm{token}}\left(m/z, I\right) = \mathrm{FF}\left(\mathrm{TE}\left(m/z\right) \parallel I\right), \tag{3}$$

where $\parallel$ denotes concatenation. A diagram of $\mathrm{PE}_{\mathrm{token}}$ is shown in Figure 1a.

2.1.2 SINUSOIDAL m/z PEAK EMBEDDING

We now consider a numerical peak embedding that utilizes the full m/z precision. We employ a sinusoidal embedding as follows:

$$\mathrm{SE}\left(m/z, 2i; d\right) = \sin\left(2\pi\left[\lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2i/(d-2)}\right]^{-1} m/z\right) \tag{4}$$

$$\mathrm{SE}\left(m/z, 2i+1; d\right) = \cos\left(2\pi\left[\lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2i/(d-2)}\right]^{-1} m/z\right). \tag{5}$$

The frequencies are chosen so that the wavelengths are log-spaced from $\lambda_{\min} = 10^{-2.5}$ Daltons to $\lambda_{\max} = 10^{3.3}$ Daltons, corresponding to the mass scales we wish to resolve. This embedding is inspired by Li et al. (2021a); Vaswani et al. (2017). Similarly to Equation 3, we construct the sinusoidal peak embedding,

$$\mathrm{PE}_{\mathrm{sin}}\left(m/z, I\right) = \mathrm{FF}\left(\mathrm{FF}\left(\mathrm{SE}\left(m/z\right)\right) \parallel I\right). \tag{6}$$

We also experiment with casting the m/z inputs as half, single, and double precision floating point values. Equations 4-5 are computed at the precision specified by m/z and the results are then cast to the precision of the model weights used for training. A diagram of $\mathrm{PE}_{\mathrm{sin}}$ is shown in Figure 1b.

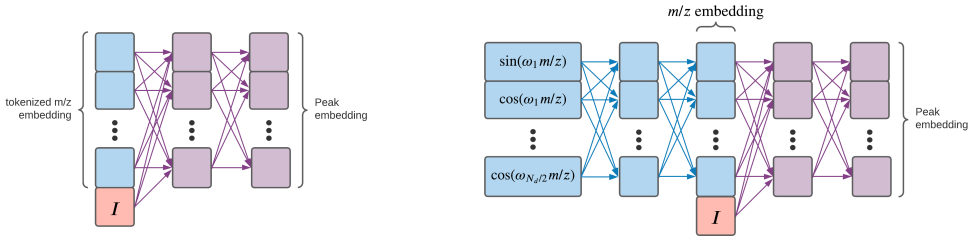
2.1.3 TRANSFORMER

The set of fragment peaks and the precursor are passed through either of the two peak embeddings specified above. The result is a sequence of embedding vectors which is then passed through a transformer encoder:

$$\mathrm{SpectrumEncoder}\left(S\right) = \mathrm{TransformerEncoder}\left(\mathrm{PE}\left(m/z, I\right)_{\mathrm{precursor}}, \ldots, \mathrm{PE}\left(m/z, I\right)_{\mathrm{fragment}_N}\right). \tag{7}$$

Here PE stands in for either $\mathrm{PE}_{\mathrm{token}}$ or $\mathrm{PE}_{\mathrm{sin}}$. TransformerEncoder has embedding dimension $d = 512$ and six layers, each with 32 attention heads and an inner hidden dimension of d. Note that all hidden and embedding layers in our peak embeddings above have dimension d as well. As mass spectra have no intrinsic ordering, we opt to not include a positional encoding.

In order to get a single embedding vector as an output, the final transformer layer query only attends to the first embedding, corresponding to the position of the precursor m/z. A diagram of our full model architecture is shown in Figure 2.



(a) Architecture of $\mathrm{PE}_{\mathrm{token}}$.

(b) Architecture of $\mathrm{PE}_{\mathrm{sin}}$.

Figure 1: Architectures used to embed MS2 peaks (composed of a m/z and an intensity I). Both approaches produce an embedding of m/z of dimension d. The m/z embedding in Figure 1a (left) is produced by a learned embedding layer. The m/z embedding in Figure 1b (right) is produced by a sinusoidal embedding followed by a two layer MLP. In both cases, an intensity value I is appended to make a vector of dimension $d + 1$. This is then fed into a simple feed-forward network to produce a peak embedding of dimension d. We use blue to highlight the network components involved in producing an m/z embedding and purple to highlight the full peak embedding after intensity is incorporated.

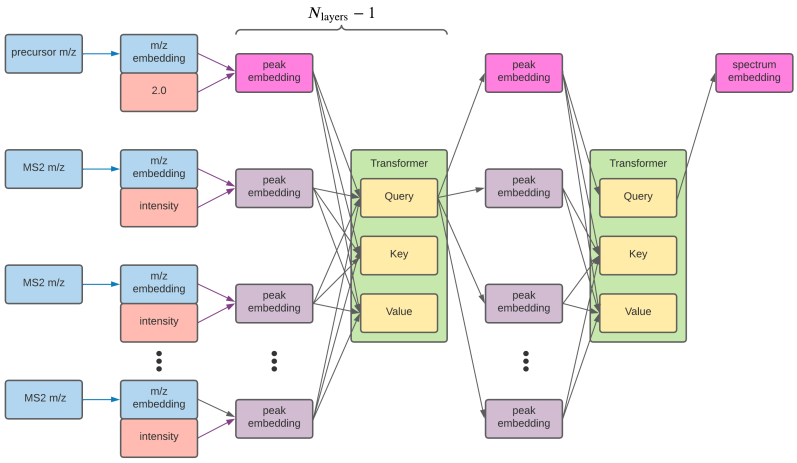


Figure 2: Full model architecture that highlights how we featurize and embed an MS2 spectrum in a dense vector space.

2.2 MODEL TRAINING

We train the base models discussed here via two independent tasks aimed at applications within metabolomics and medicinal chemistry. These tasks are described in detail below. We build these models using PyTorch (Falcon & The PyTorch Lightning team, 2019; Paszke et al., 2019) and train them with the Adam (Kingma & Ba, 2015) optimization algorithm with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate $\alpha = 5.0 \times 10^{-5}$. We also use a weight decay (Krogh & Hertz, 1991) parameter of 0.1, a gradient clipping a parameter of 0.5, and a dropout (Srivastava et al., 2014) parameter of 0.1. All models were trained using between 25 and 50 epochs. Finally, all model weights were trained using half precision floating point values, regardless of the precision of the m/z inputs.

2.2.1 SPECTRAL SIMILARITY WITH SIAMESE NETWORKS

A common workflow for the analysis of mass spectrometry data is to compute some notion of spectral similarity for pairs of spectra that is intended to correlate with the molecular similarity of the underlying compounds (Yilmaz et al., 2016). This can then be used for tasks such as spectral library search (Stein & Scott, 1994) and molecular networking (Quinn et al., 2017; Watrous et al., 2012). Historically, most spectral similarity metrics used in the field have been heuristic-based (Li et al., 2021b; Yilmaz et al., 2016). More recently, spectral similarity methods that make use of techniques from deep learning have been developed (Huber et al., 2021a;b). The latter of these, MS2Deepscore, is a Siamese network that trains a simple MLP (by binning m/z) to predict a measure of molecular similarity (Bajusz et al., 2015) from pairs of mass spectra. The loss function is the MSE between the ground truth molecular similarity and the cosine similarity evaluated on the dense embedding vectors generated by the base model.

Molecular similarity scores range from 0 to 1. The scores of randomly sampled pairs from our labeled dataset are not uniformly distributed and strong similarity matches are rare. Huber et al. (2021b), describe a procedure to sample pairs of spectra such that the molecular similarities will be uniformly distributed. We use this procedure to sample pairs for training and evaluation tasks.

Following Huber et al. (2021b) we also train a Siamese network, but we do so using the base model transformer architectures outlined in Section 2.1 instead of with the binned mass vectors used by Huber. We report performance results for models trained using $\mathrm{PE}_{\mathrm{token}}$ and $\mathrm{PE}_{\mathrm{sin}}$ below in Section 3.2.

2.2.2 PROPERTY PREDICTION

We also train the base model architecture to predict 10 chemical properties relevant to medicinal chemistry directly from an MS2 spectrum. We do so by training the following network,

$$\mathrm{Properties} = \mathrm{FF}\left(\mathrm{SpectrumEncoder}\left(S\right)\right), \tag{8}$$

where the final layer has dimension equal to the number of properties we wish to predict. We also scale all of our training labels to zero mean and unit variance. For inference, we run predicted properties through an inverse-scaler. As in Section 2.2.1, we train multiple versions of this property prediction model. Due to the lack of existing methods for comparison, we also train a version of Equation 8 on binned spectrum representations for a simple baseline. We replace $\mathrm{SpectrumEncoder}$ with the MS2Deepscore feed forward architecture found in Huber et al. (2021b) for this baseline.

The properties we predict are standard indicators of druglikeness and bioavailability in the field of medicinal chemistry. These include the properties that make up the Quantitative Estimate of Druglikeness (QED) along with several others (Bickerton et al., 2012; Lipinski et al., 2001; Veber et al., 2002). For the complete list of properties, see Table 2. All properties are computed deterministically from chemical structure using RDKit (Landrum, 2016), so there is no additional dependency on experimental data.

3 RESULTS

3.1 DATA

We construct a set of labeled MS2 spectra by combining a number of publicly available datasets (nis, 2020; Aoyagi et al., 2017; Han et al., 2021; Horai et al., 2010a;b; Koelmel et al., 2020; Lee et al., 2020; Lei et al., 2015; Mardal et al., 2019; Phapale et al., 2021; Sawada et al., 2012; Smith et al., 2005; Tada et al., 2019; Wang et al., 2016) with our internal proprietary dataset. We allow spectra collected in both positive and negative ion mode, and across a variety of instrument types, collision energies, and other important instrument parameters. We constrain that spectra in our dataset have at least 5 peaks and at least 3 decimal places of m/z resolution. We strip all stereochemistry from our molecular structure labels, which is a common step taken in MS2 modeling and allows for better molecule-disjoint splitting. Chemical properties and fingerprints are then computed from the cleaned molecules using RDKit (Landrum, 2016). Our resulting dataset has $1{,}251{,}830$ spectra corresponding to $45{,}351$ distinct molecules.

In metabolomics experiments, biological samples contain both compounds that have previously been profiled in MS2 libraries and compounds that have not been profiled. Therefore, we separately characterize performance on "known" molecules that have been previously profiled with MS2, and "novel" molecules for which no labeled MS2 spectra are available. We partition our set of labeled spectra into a training, development and test set of size $1{,}214{,}812$, $18{,}750$, and $18{,}268$ respectively. All spectra associated with $1{,}002$ randomly selected molecules are partitioned together into each of the development and test sets. In this manner, a subset of both the development and test sets is molecule-disjoint from the training set. Henceforth, we refer to any spectrum in the development or test set as corresponding to a known molecule if that molecule is present in the training set, and a novel molecule if it is not. Our resulting development and test sets each have 998 known spectra and $17{,}752$ and $17{,}270$ novel spectra, respectively. While a subset of molecules in our development and test sets overlap with our training set, the spectra in all three sets are fully disjoint. We use the development set for hyperparameter optimization and only access the test set to compute reported metrics.

3.2 SPECTRAL SIMILARITY

A key application of any spectral similarity method is spectral library search (Stein & Scott, 1994) (or dereplication), whereby an unknown MS2 spectrum is identified by searching through a database of curated spectra obtained from known molecules. To benchmark our various spectral similarity models, we query spectra from our test set against a lookup reference (which is the same as our training set). Every query spectrum is matched, by maximum similarity, to a reference spectrum. We then assess accuracy on a set of spectra queries by determining the fraction of matches that exceed a particular molecular similarity threshold. We use two such thresholds, 0.60 for approximate match and 0.95 for nearly exact match. Finally, we typically have a varying number of recorded spectra per distinct molecule so all reported accuracies are macro averaged over query molecules.

Because novel molecules do not occur in the reference set, we estimate the maximum theoretical performance on the prediction task for novel molecules by evaluating the best molecular similarity matches between the novel test set and the reference set. A spectral similarity model that perfectly approximates molecular similarity would score 0.31 and 0.89 on nearly exact and approximate matching, respectively. We also evaluate the performance of a random baseline by uniformly sampling pairs of spectra and computing their corresponding molecular similarity. Only 0.713% and 0.03% of random spectra pairs from our data have a corresponding molecular similarity above the threshold for an approximate and nearly exact match, respectively.

We train a number of spectral similarity models, as described in Section 2.2.1, and will refer to the sinusoidal Siamese transformer and tokenized Siamese transformer models as the sinusoidal and tokenized models respectively. For baselines, we train both a Spec2Vec (Huber et al., 2021a) and an MS2Deepscore (Huber et al., 2021b) model on our data. In Table 1 we report spectral library search accuracies (based on molecular similarity thresholds) for both known and novel compounds obtained from a number of spectral similarity models. In the same table we also report the MSE loss between predicted and actual molecular similarity on spectrum pairs drawn from train, known,

Table 1: Spectral library search performance

Spectral Set	MSE			Spectral library search accuracy			
	train	known	novel	known		novel	
Threshold				0.95	0.60	0.95	0.60
Spec2Vec				0.911	0.954	0.078	0.414
MS2Deepscore	0.026	0.029	0.047	0.865	0.918	0.072	0.387
Tokenized m/z	0.017	0.019	0.030	0.94	0.961	0.064	0.382
Sinusoidal m/z (`float16`)	0.019	0.019	0.03	0.918	0.955	0.084	0.401
Sinusoidal m/z (`float64`)	**0.013**	**0.015**	**0.024**	**0.943**	**0.97**	**0.096**	**0.435**

and novel spectra sets. No MSE can be reported for Spec2Vec as this method does not predict a molecular similarity.

We find that Spec2Vec has strong, only surpassed by the double precision sinusoidal model, spectral library search performance on novel, but only beats MS2Deepscore on known. MS2Deepscore only outperforms the tokenized model on novel and underperforms everything on known. Moreover, MS2Deepscore is substantially less accurate at predicting the specific pair molecular similarity on known and novel spectra when compared to all transformer approaches. The double precision sinusoidal model produced the best performance on all evaluation tasks, beating all baselines. This model produces state of the art known spectral library search accuracies of 0.943 and 0.97 for nearly exact and approximate matching respectively. Numerical representation of MS2 m/z via sinusoidal embeddings consistently outperforms all the discretization based approaches we benchmarked.

To assess whether the multi-scale embeddings are learning from high resolution information, we also train and evaluate versions of our sinusoidal model where m/z's are cast to half precision floating point values. Half precision floating point can only resolve up to approximately parts per $\mathcal{O}(10,000)$, far below the parts per million precision of experimental mass spectrometry data. The performance of the half precision model, reported in Table 1, drops to be on par with the tokenized model (worse on known and better on novel). We do not report results with single precision m/z since these were indistinguishable from double precision m/z. On all evaluation metrics, using m/z inputs of precision greater than 16 bits dramatically improves performance of the sinusoidal model. From this, we draw the implication that sinusoidal embedding of m/z is able to effectively learn from with the wide range of scales present in high precision mass spectrometry measurements and this partially accounts for the superior performance of this method.

3.2.1 Qualitative embedding analysis

To further characterize the respective properties of token and multi-scale sinusoidal embeddings, we inspect UMAP (McInnes et al., 2018) projections of our m/z embeddings in Figure 3. For this analysis, we use siamese transformer models described in Section 2.2.1. For our sinusoidal models, we embed $50{,}000$ m/z values between 0 and $1{,}000\,\mathrm{Daltons}$ using $\mathrm{FF}\left(\mathrm{SE}\left(m/z\right)\right)$. We do not use the full peak embedding function as we are not interested in the embedding of intensity information. Because our tokenization procedure involves rounding m/z values to 1 decimal place, for the tokenized model we only embed $10{,}000$ m/z values between 0 and $1{,}000\,\mathrm{Daltons}$ using the embedding function $\mathrm{TE}\left(m/z\right)$.

While m/z measurements are continuous, the space they represent is inherently discrete. Two molecular fragments differing in atomic composition (let alone structure) could nevertheless have a very similar m/z's. However, the large separation between molecular mass scales and the mass scale at which nuclear forces manifest themselves means that the relative atomic composition may be deduced from the fractional m/z (Jones et al., 2004; Pourshahian & Limbach, 2008), defined by,

$$\{m/z\} = m/z - \lfloor m/z \rfloor \,. \tag{9}$$

Therefore, we expect a quality embedding of m/z to embed fragments similar in both m/z and $\{m/z\}$ close to one another, while paying greater attention to the latter.

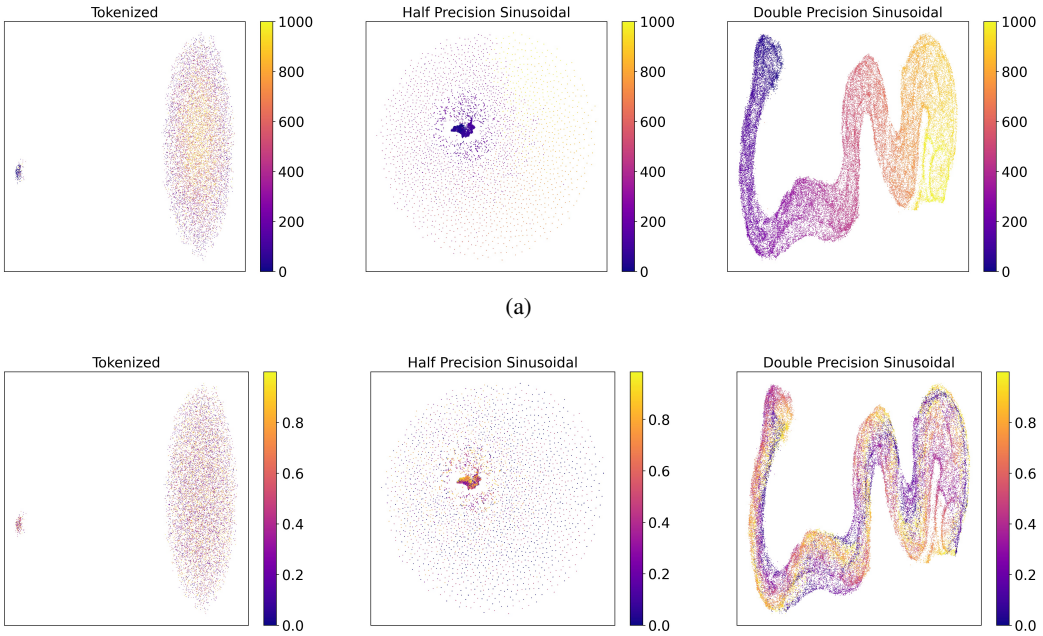


Figure 3: (3a) UMAP projections of m/z embeddings colored by m/z value. (3b) UMAP projections of m/z embeddings colored by $\{m/z\}$. High resolution sinusoidal embeddings imbue model latent space with additional structure not found in other baseline models.

As is seen in Figure 3, embeddings from tokenized and low precision sinusoidal models are able to capture general trends in m/z. However, they fail to preserve distances in m/z and show little to no structure in $\{m/z\}$. In contrast, the high precision sinusoidal embeddings allow our models to represent important information in $\{m/z\}$, while preserving distance in m/z.

3.3 Property Prediction

Many applications in metabolomics can be unblocked with knowledge of just a limited set of chemical properties, without need for identification of molecular structure. Towards this end, we investigate the parallel task of chemical property prediction from MS2 spectra. We evaluate sinusoidal *vs*. tokenized m/z embedding performance on a list of properties that have compelling applications in drug discovery and are easily computable from our molecule labels. For this experiment, we train a sinusoidal model and several baseline models as described in Section 2.2.2.

In Table 2, property prediction models are evaluated on known and novel structures using R^2. All transformer models outperform our feed forward baseline, illustrating the difficulty of the task and the significance of our results. Half precision sinusoidal models produced an R^2 of 0.746 (0.936) on known (novel) molecules averaged across all 10 properties. These results are competitive with tokenization but substantially underperform when compared to higher precision sinusoidal embeddings. When m/z values are specified at double precision, sinusoidal embeddings show improvements over tokenization on each property, resulting in an 11% improvement over tokenization on R^2 for novel molecules averaged across all properties.

This is further evidence that the multi-scale information captured by sinusoidal embeddings allows models to better understand the nuances of MS2 and chemical space. In addition, the substantial performance of our sinusoidal property prediction model is strong enough to enable prioritization over a large number of unknown metabolites. This has major implications in the drug discovery context, as inspecting the properties of unknown molecules in complex mixtures was previously not possible without isolation and individual experimentation.

Table 2: R^2 for predicted chemical properties

Property	Feed Forward		Tokenized m/z		Sinusoidal m/z	
	Known	Novel	Known	Novel	Known	Novel
all	0.824	0.604	0.943	0.72	**0.976**	**0.8**
atomic $\log P$	0.768	0.357	0.919	0.437	**0.968**	**0.622**
number of hydrogen bond acceptors	0.835	0.762	0.969	0.887	**0.987**	**0.941**
number of hydrogen bond donors	0.834	0.572	0.924	0.683	**0.968**	**0.69**
polar surface area	0.851	0.716	0.965	0.862	**0.986**	**0.907**
number of rotatable bonds	0.801	0.669	0.939	0.792	**0.976**	**0.838**
number of aromatic rings	0.838	0.372	0.934	0.474	**0.981**	**0.655**
number of aliphatic rings	0.839	0.638	0.943	0.762	**0.974**	**0.821**
number of heteroatoms	0.823	0.719	0.968	0.902	**0.988**	**0.946**
fraction of sp3 carbons	0.846	0.618	0.951	0.723	**0.977**	**0.821**
quantitative estimate of druglikeness	0.806	0.617	0.912	0.681	**0.952**	**0.755**

4 Conclusions

Through a number of experiments, we demonstrate that MS2 modeling methods that use a numerical representation of m/z via multi-scale sinusoidal embeddings are superior to methods which make use of discretized m/z. In order to obtain insight into the performance attributed to sinusoidal embeddings, we present evidence that indicates that sinusoidal embeddings enable learning from the high-precision data that is produced from MS2 experiments.

The modeling work represented here produces state of the art results on spectral dereplication tasks. We also produce strong results on ML chemical property prediction from MS2 spectra. The property prediction results of are sufficient quality to functionally inform and advance drug-discovery efforts. We expect that sinusoidal embeddings of m/z will be a critical component of further research that incorporates MS2 spectra modeling, and we look forward to exploring their use in other contexts.

Reproducibility Statement

The source of all of the data used in this work and necessary preprocessing steps are outelined in Section 3.1. While we cannot share our proprietary souce code, the model architecture is explained in detail in Section 2. Finally, all parameters used in model training are specified in Section 2.2.

References

Nist standard reference database 1a. https://www.nist.gov/srd/nist-standard-reference-database-1a, Jun 2020.

Ryohei Aoyagi, Kazutaka Ikeda, Yosuke Isobe, and Makoto Arita. Comprehensive analyses of oxidized phospholipids using a measured MS/MS spectra library. *Journal of Lipid Research*, 58 (11):2229–2237, November 2017. doi: 10.1194/jlr.d077123. URL https://doi.org/10.1194/jlr.d077123.

David Bajusz, Anita Racz, and Karoly Heberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015. doi: 10.1186/s13321-015-0069-3.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, January 2012. doi: 10.1038/nchem.1243. URL https://doi.org/10.1038/nchem.1243.

Sebastian Böcker and Florian Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24(16):i49–i55, 08 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn270. URL https://doi.org/10.1093/bioinformatics/btn270.

Warwick B. Dunn, Alexander Erban, Ralf J.M. Weber, Darren J. Creek, Marie Brown, Rainer Breitling, Thomas Hankemeier, Royston Goodacre, Steffen Neumann, Joachim Kopka, and Mark R. Viant. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9(1):S44–S66, 2013. ISSN 1573-3882. doi: 10.1007/s11306-012-0434-4. Relation: http://www.rug.nl/research/gbb/ Rights: University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute.

Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015. doi: 10.1073/pnas.1509788112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1509788112.

William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. URL https://github.com/PyTorchLightning/pytorch-lightning.

Gary L. Glish and Richard W. Vachet. The basics of mass spectrometry in the twenty-first century. *Nature Reviews Drug Discovery*, 2(2):140–150, February 2003. doi: 10.1038/nrd1011. URL https://doi.org/10.1038/nrd1011.

Shuo Han, Will Van Treuren, Curt R. Fischer, Bryan D. Merrill, Brian C. DeFelice, Juan M. Sanchez, Steven K. Higginbottom, Leah Guthrie, Lalla A. Fall, Dylan Dodd, Michael A. Fischbach, and Justin L. Sonnenburg. A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature*, 595(7867):415–420, July 2021. doi: 10.1038/s41586-021-03707-9. URL https://doi.org/10.1038/s41586-021-03707-9.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, July 2010a. doi: 10.1002/jms.1777. URL https://doi.org/10.1002/jms.1777.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010b. doi: https://doi.org/10.1002/jms.1777. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jms.1777.

Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):1–18, 02 2021a. doi: 10.1371/journal.pcbi.1008724. URL https://doi.org/10.1371/journal.pcbi.1008724.

Florian Huber, Sven A. van der Burg, Justin J. J. van der Hooft, and Lars Ridder. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics*, 13, 2021b.

Jeffrey J. Jones, Michael J. Stump, Richard C. Fleming, Jackson O. Lay, and Charles L. Wilkins. Strategies and data analysis techniques for lipid and phospholipid chemistry elucidation by intact cell MALDI-FTMS. *Journal of the American Society for Mass Spectrometry*, 15(11):1665–1674, 2004. ISSN 1044-0305. doi: 10.1016/j.jasms.2004.08.007.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. *ICLR 2015*, 2015.

Jeremy P. Koelmel, Matthew K. Paige, Juan J. Aristizabal-Henao, Nicole M. Robey, Sara L. Nason, Paul J. Stelben, Yang Li, Nicholas M. Kroeger, Michael P. Napolitano, Tina Savvaides, Vasilis Vasiliou, Pawel Rostkowski, Timothy J. Garrett, Elizabeth Lin, Chris Deigl, Karl Jobst, Timothy G. Townsend, Krystal J. Godri Pollitt, and John A. Bowden. Toward comprehensive per- and polyfluoroalkyl substances annotation using FluoroMatch software and intelligent high-resolution tandem mass spectrometry acquisition. *Analytical Chemistry*, 92(16):11186–11194, August 2020. doi: 10.1021/acs.analchem.0c01591. URL https://doi.org/10.1021/acs.analchem.0c01591.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *NIPS*, 1991.

Svetlana Kutuzova, Christian Igel, Mads Nielsen, and Douglas McCloskey. Bi-modal variational autoencoders for metabolite identification using tandem mass spectrometry. *bioRxiv*, 2021. doi: 10.1101/2021.08.03.454944. URL https://www.biorxiv.org/content/early/2021/08/04/2021.08.03.454944.

Greg Landrum. Rdkit: Open-source cheminformatics software. 2016. URL https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

Sangwon Lee, Sungbo Hwang, Myungwon Seo, Ki Beom Shin, Kwang Hoe Kim, Gun Wook Park, Jin Young Kim, Jong Shin Yoo, and Kyoung Tai No. BMDMS-NP: A comprehensive ESI-MS/MS spectral library of natural compounds. *Phytochemistry*, 177:112427, September 2020. doi: 10.1016/j.phytochem.2020.112427. URL https://doi.org/10.1016/j.phytochem.2020.112427.

Zhentian Lei, Li Jing, Feng Qiu, Hua Zhang, David Huhman, Zhiqin Zhou, and Lloyd W. Sumner. Construction of an ultrahigh pressure liquid chromatography-tandem mass spectral library of plant natural products and comparative spectral analyses. *Analytical Chemistry*, 87(14):7373–7381, July 2015. doi: 10.1021/acs.analchem.5b01559. URL https://doi.org/10.1021/acs.analchem.5b01559.

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding. *arXiv*, 2021a.

Yuanyue Li, Tobias Kind, Jacob Folz, Arpana Vaniya, Sajjan Singh Mehta, and Oliver Fiehn. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods*, pp. 1–8, 2021b. ISSN 1548-7091. doi: 10.1038/s41592-021-01331-z.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1pii of original article: S0169-409x(96)00423-1. the article was originally published in advanced drug delivery reviews 23 (1997) 3-25.1. *Advanced Drug Delivery Reviews*, 46(1):3–26, 2001. ISSN 0169-409X. doi: https://doi.org/10.1016/S0169-409X(00)00129-0. URL https://www.sciencedirect.com/science/article/pii/S0169409X00001290. Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998.

Eleni Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia Kavraki. Spec2mol: An end-to-end deep learning framework for translating ms/ms spectra to de-novo molecules. *ChemRxiv*, 2021. doi: 10.26434/chemrxiv-2021-6rdh6.

Marie Mardal, Mette Findal Andreasen, Christian Brinch Mollerup, Peter Stockham, Rasmus Telving, Nikolaos S Thomaidis, Konstantina S Diamanti, Kristian Linnet, and Petur Weihe Dalsgaard. HighResNPS.com: An online crowd-sourced HR-MS database for suspect and non-targeted screening of new psychoactive substances. *Journal of Analytical Toxicology*, 43(7):520–527, May 2019. doi: 10.1093/jat/bkz030. URL https://doi.org/10.1093/jat/bkz030.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL http://arxiv.org/abs/1802.03426. cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Prasad Phapale, Andrew Palmer, Rose Muthoni Gathungu, Dipali Kale, Britta Brügger, and Theodore Alexandrov. Public LC-orbitrap tandem mass spectral library for metabolite identification. *Journal of Proteome Research*, 20(4):2089–2097, February 2021. doi: 10.1021/acs.jproteome.0c00930. URL https://doi.org/10.1021/acs.jproteome.0c00930.

Soheil Pourshahian and Patrick A. Limbach. Application of fractional mass for the identification of peptide-oligonucleotide cross-links by mass spectrometry. *Journal of Mass Spectrometry*, 43(8): 1081–1088, 2008. ISSN 1096-9888. doi: 10.1002/jms.1391.

Rui Qiao, Ngoc Hieu Tran, Lei Xin, Baozhen Shan, Ming Li, and Ali Ghodsi. DeepNovoV2: Better de novo peptide sequencing with deep learning. *arXiv*, 2019.

Robert A. Quinn, Louis-Felix Nothias, Oliver Vining, Michael Meehan, Eduardo Esquenazi, and Pieter C. Dorrestein. Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends in Pharmacological Sciences*, 38(2):143–154, 2017. ISSN 0165-6147. doi: 10.1016/j.tips.2016.10.011.

Yuji Sawada, Ryo Nakabayashi, Yutaka Yamada, Makoto Suzuki, Muneo Sato, Akane Sakata, Kenji Akiyama, Tetsuya Sakurai, Fumio Matsuda, Toshio Aoki, Masami Yokota Hirai, and Kazuki Saito. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. *Phytochemistry*, 82:38–45, October 2012. doi: 10.1016/j.phytochem.2012.07.007. URL https://doi.org/10.1016/j.phytochem.2012.07.007.

Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B. Kell. MassGenie: a transformer-based deep learning method for identifying small molecules from their mass spectra. *bioRxiv*, pp. 2021.06.25.449969, 2021. doi: 10.1101/2021.06.25.449969.

Colin A Smith, Grace O?? Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. METLIN. *Therapeutic Drug Monitoring*, 27(6):747–751, December 2005. doi: 10.1097/01.ftd.0000179845.53213.39. URL https://doi.org/10.1097/01.ftd.0000179845.53213.39.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 2014.

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

Ipputa Tada, Hiroshi Tsugawa, Isabel Meister, Pei Zhang, Rie Shu, Riho Katsumi, Craig E. Wheelock, Masanori Arita, and Romanas Chaleckis. Creating a reliable mass spectral–retention time library for all ion fragmentation-based metabolomics. *Metabolites*, 9(11):251, October 2019. doi: 10.3390/metabo9110251. URL https://doi.org/10.3390/metabo9110251.

Justin Johan Jozias van der Hooft, Joe Wandy, Michael P. Barrett, Karl E. V. Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743, 2016. doi: 10.1073/pnas.1608041113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1608041113.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv*, 2017.

Daniel F. Veber, Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623, 2002. doi: 10.1021/jm020017n. URL https://doi.org/10.1021/jm020017n. PMID: 12036371.

Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard Ø Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8):828–837, August 2016. doi: 10.1038/nbt.3597. URL https://doi.org/10.1038/nbt.3597.

Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S. Heath, Jane Y. Yang, Roland D. Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M. Raaijmakers, Bradley S. Moore, Julia Laskin, Nuno Bandeira, and Pieter C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, 109(26):E1743–E1752, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1203689109.

Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Sewoong Oh, and William Stafford Noble. De novo mass spectrometry peptide sequencing with a transformer model. *bioRxiv*, pp. 2022.02.07.479481, 2022. doi: 10.1101/2022.02.07.479481.

Şule Yilmaz, Elien Vandermarliere, and Lennart Martens. Methods to Calculate Spectrum Similarity. *Proteome Bioinformatics*, pp. 75–100, 2016. ISSN 1064-3745. doi: 10.1007/978-1-4939-6740-7_7.

Xi-wu Zhang, Qiu-han Li, Zuo-di Xu, and Jin-jin Dou. Mass spectrometry-based metabolomics in health and medical science: a systematic review. *RSC Adv.*, 10:3092–3104, 2020. doi: 10.1039/C9RA08985C. URL http://dx.doi.org/10.1039/C9RA08985C.