nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00708-3

Annotating metabolite mass spectra with domain-inspired chemical formula transformers

In the format provided by the authors and unedited

S1 Supplementary Text

S1.1 Additional model details

Unfolding layers. Novel unfolding modules are introduced in MIST to progressively grow the fingerprint prediction. Target fingerprints are compressed using modulo operations to derive "lower resolution" targets. These lower dimensional vectors inherently have collisions for properties. That is, rather than each bit indicating the presence of a single property or motif, each bit represents the presence of one of several properties (Fig. S1).

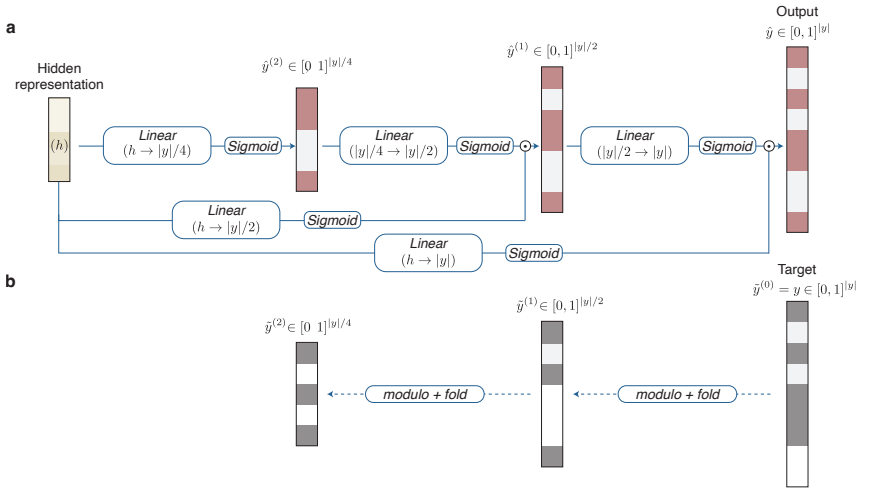

Fig. S1 Unfolding fingerprint prediction module. a. The module starts with a hidden representation and subsequently predicts increasing resolutions of the fingerprint. **b.** At each intermediate fingerprint prediction $\hat{y}^{(i)}$, a corresponding target vector is generated by "folding" the previous, higher resolution fingerprint. An intermediate loss is calculated at each of these resolutions.

Binned feed forward baselines. A simple feed forward network inspired by MetFID is utilized as a baseline (Fig. S2) [1].

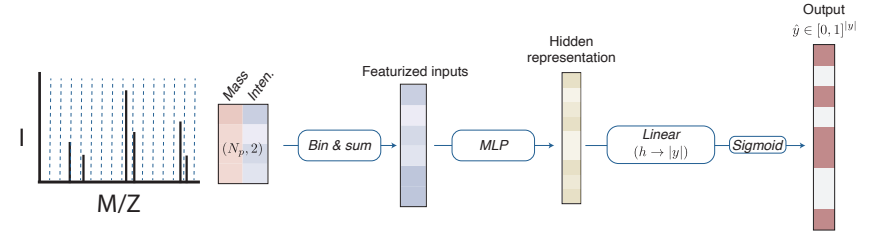

Fig. S2 Feed forward network binned spectra baseline. An input spectrum is binned, a multilayer perceptron (MLP) is applied, and the output is projected to a fingerprint prediction.

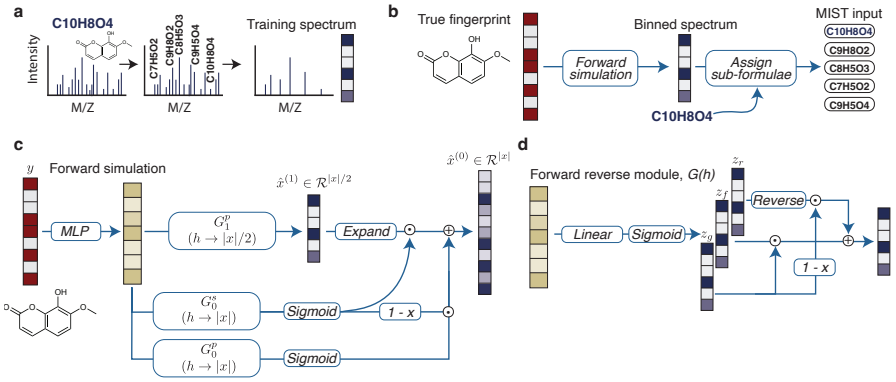

Fig. S3 Forward simulation of spectra from molecules. a. An input spectrum, molecule pair is first denoised and cleaned by labeling chemical sub-formulae. This is converted into a binned spectra representation. **b.** To generate new model training examples, molecules are fingerprinted, binned spectra predictions are predicted by a neural network, and sub-formula are assigned to each of the bins such that the spectra can be used as input to MIST. **c.** The forward simulation neural architecture utilizes unfolding layers. A multilayer perceptron (MLP) projects the molecular fingerprint into a hidden representational space. Forward-reverse models, G, are used to progressively grow the binned prediction such that resolution increases at each step, similar to unfolding layers. **d.** Illustration of the forward-reverse module, where intensities (z_f) and neutral losses (z_r) are predicted. Neutral losses are reversed and mapped onto intensity bins using the full mass as input then summed according to a learned gate to get a single prediction of intensities in each bin.

S1.2 CSI:FingerID annotations for Mills et al. cohort

In Section 2.5, we conduct a prospective analysis to showcase how MIST can be used with real world data. To evaluate the reasonableness of our annotations against the existing tool, SIRIUS and CSI:FingerID [2], we repeat the annotation process with these tools. Concretely, we use an equivalent MGF

spectrum extraction process and pass this into SIRIUS Version 5.6.3 for end-to-end formula, fingerprint, and structure annotation. We use SIRIUS Version 5.6.3 as SIRIUS Version 4.9 was deprecated during the course of this work.

We constrain the compound search to the HMDB database [3] to match our MIST pipeline. SIRIUS 5 stalled out and failed on the full MGF file including larger compounds; correspondence with the authors revealed that larger fragmentation trees can lead to memory errors. As such, we subsetted the dataset to feature only the 1,515 spectra with precursor ions under 500 Da. 545 compounds were annotated with structures, of which 342 had equivalent chemical formulae to those we extracted with SIRIUS 4.9 earlier. From this set, MIST predicted equivalent InChIKeys on 58% of the spectra; a random selection from HMDB isomers matched SIRIUS on 32%. We include a full table showing drop-off at each step in the processing pipeline (Table S1).

Surprisingly, we observe in this workflow that the chemical formulae of compounds predicted by SIRIUS are different from chemical formulae we exported from SIRIUS as inputs to MIST. This is due to SIRIUS *re-ranking* chemical formulae during the compound retrieval phase of their pipeline. Testing this, we run SIRIUS again with PubChem as a database retrieval and find that of the 1,355 compounds that are annotated with chemical formulae in the initial step, only 625 of these have the same chemical formulae after compound retrieval.

To probe the robustness of our findings at the chemical class level, we repeat the MIST processing pipeline with two changes: (1) we utilize 640 spectra and formula annotations from the final step of SIRIUS's pipeline, after re-ranking during retrieval from PubChem and (2) utilize PubChem as a retrieval database. Under this analysis, we find 162 of the 640 spectra have equivalent

annotations under SIRIUS and MIST, a greater than 25% agreement even when searching the much larger PubChem database.

Using the resulting 640 revised annotations outputted from MIST, we repeat our prospective IBD analysis pipeline and find that dipeptides are still one of the most correlated classes with disease severity (Fig. S4). Further, when looking for differential abundance, we see that piperidine and pyridine alkaloids, are still differentially abundant in CD patients (Fig. S5). These class level differences are no longer present for UC patients. This discordance emphasizes that automated annotation pipelines are not yet robust to changes in processing parameters such as spectrum preprocessing, formula annotations, fingerprint prediction, and retrieval library selection.

In the future, moving beyond fragmentation tree based predictions for formula annotation will enable efficient annotation of larger compounds. Improving automated chemical formula annotation and database retrieval library construction will similarly lead to higher quality and more consistent annotations.

Table S1 Comparison of SIRIUS and MIST for annotating IBD cohort data.

	Total spectra
MIST Pipeline	1,990
SIRIUS Pipeline (< 500 Da)	1,515
MIST identifies a structure in HMDB	705
SIRIUS identifies a structure in HMDB	724
SIRIUS and MIST both identify a structure in HMDB	545
SIRIUS and MIST annotate compounds with equiv. formula	342
SIRIUS and MIST find same InChIKey	200
SIRIUS and Random-HMDB find same InChIKey	110

S1.3 Dataset splits

For both the proprietary NIST data and public benchmarking data, care is taken to ensure that no two molecules with the same InChIKey appear across separate dataset split folds. We can quantify and demonstrate this effect by

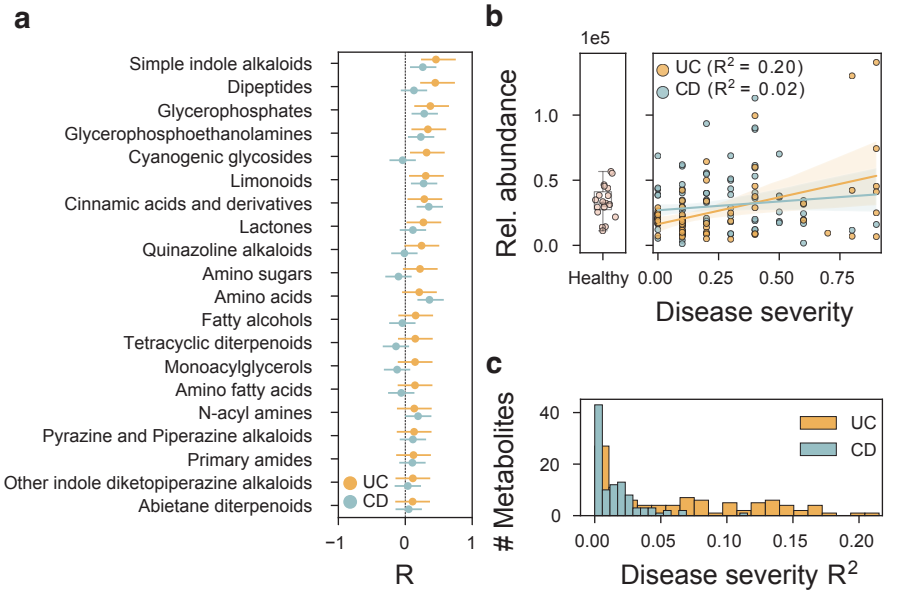

Fig. S4 Repeated analysis of clinically relevant dipeptides with the PubChem retrieval library. a. Metabolite classes ranked by their Pearson correlation with IBD disease severity. All putative metabolites are assigned disease classes with NPClassifier [4] and relative abundances are summed across classes. Within the ulcerative colitis (UC) and Crohn's disease (CD) cohorts, metabolite classes are correlated with disease severity and sorted according to their correlation with UC (top 20 shown; each computed for n=60 patients with UC, n=102 CD patients). Error bars indicate 95% confidence intervals around the mean. **b.** Dipeptide relative abundances for each patient are plotted against disease severity within the healthy (n=19), UC, and CD cohorts. Dipeptides correlate with disease severity for UC patients ($R^2 = 0.20$, $p = 0.02$, two-sided the Wald test, adjusted with Benjamini-Hochberg) as observed by Mills et al. [5]. In the left boxplot, the data median line is shown; boxes show the interquartile range, with whiskers indicating 1.5x interquartile ranges. Regression lines are shown with 95% confidence intervals computed with bootstrapping. **c.** Within the dipeptide class, the distribution of individual metabolite correlations are shown for both UC and CD cohorts. Compound annotations are made using an ensemble of 5 MIST models, contrastive distance retrieval, and PubChem reference databases of molecules. Input spectra are subsetted to be under 500 Da, have [M+H]+ adducts, and utilize chemical formula as output by CSI:FingerID at the end of the compound identification step.

computing the nearest neighbor compound (according to Tanimoto similarity) in the training set for all molecules in the dataset. By plotting the distributions of the "maximum similarity neighbor" for both datasets, we can see that the training set features equivalent molecules (i.e., a neighbor with similarity 1.0) with a distribution mean of 0.69 and 0.71 Tanimoto similarity across the private and public datasets. In comparison, the test set distributions are more

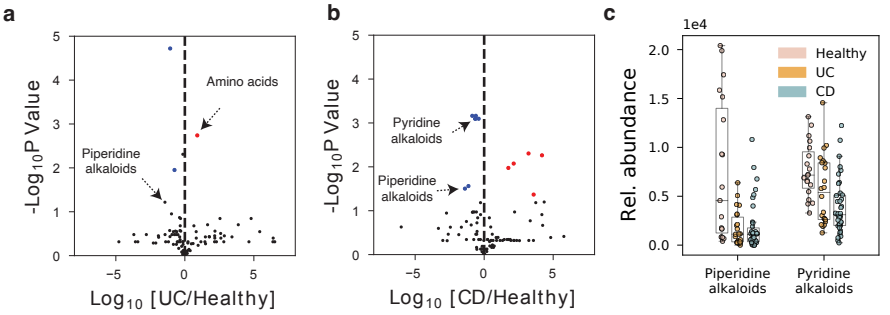

Fig. S5 Reanalyzing putative metabolite class differences between healthy and diseased IBD cohorts using the PubChem database. a,b. Putative metabolite classes from the Mills et al. cohort [5] are scattered showing their fold change and respective p value comparing and ulcerative colitis (UC) (**a**) and Crohn's disease (CD) (**b**) cohorts to the healthy cohort. P values were computed using independent two-sided t-tests and adjusted for multiple hypothesis testing with the Benjamini Hochberg method. To have equal cohort sizes for healthy and diseased patients, UC and CD cohorts were subsetted to patients with disease severity score > 0.2. Blue scattered points indicate differentially less abundant compounds or classes; red indicates differentially increased abundance. **c.** Distributions of patients' relative abundances for piperidine and pyridine alkaloid classes are shown for healthy (n=19), UC (n=22), and CD (n=38) groups. Median lines are shown; boxes show the interquartile range, with whiskers indicating 1.5x interquartile ranges. All compound annotations are made using an ensemble of 5 MIST models, contrastive distance retrieval, and the PubChem reference database of molecules. The top chemical class annotation is found by running the top putative prediction through NPClassifier [4]. Input spectra are subsetted to be under 500 Da, have [M+H]+ adducts, and utilize chemical formula as output by CSI:FingerID at the end of the compound identification step.

normal with means of 0.63 and 0.66 across the two datasets respectively (Fig. S6).

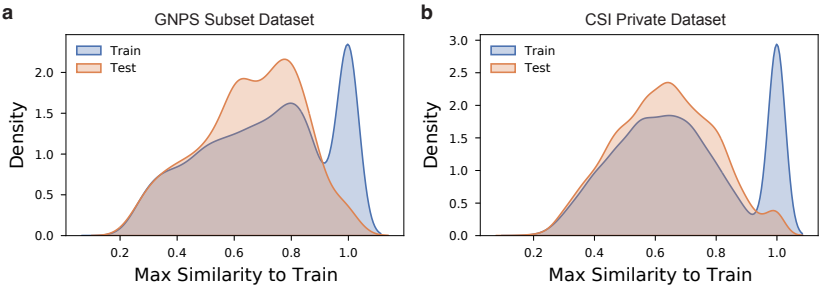

Fig. S6 Dataset split similarity. Distribution plot of the most similar molecule in the training set according to Tanimoto similarity for molecules in the train and molecules in the test sets. Distributions are computed on the public GNPS CANOPUS dataset (**a.**) and private proprietary CSI dataset (**b.**).

S1.4 Unfolding layer ablation

MIST contains unfolding layers for fingerprint prediction. We find that this leads to improved performance via an ablation study. In our ablation study, we replace the unfolding module with a single linear layer output to predict fingerprints. To verify that the unfolding mechanism (rather than just more layers) is driving performance, we add a shallow 3 layer FFN to the top of the *MIST - unfolding* model to match the 4 unfolding layers in the Unfolding module, yielding a total of 15.1M parameters vs. 3.1M parameters for *MIST* and 2.8M parameters for *MIST - unfolding*. We differentiate the two MIST ablation models as "MIST - unfolding (linear)" and "MIST - unfolding (FFN)" respectively. To avoid high computational cost of additional model features, we train these models without the auxiliary MAGMa loss and without simulated spectra. As can be seen in Table S2, we find that without unfolding, the linear and FFN output models are capable of achieving 0.6701 and 0.6617 cosine similarity vs. 0.6818 from the unfolding model. This validates the improvements offered by the unfolding layer.

Table S2 Fingerprint prediction accuracy on CANOPUS GNPS subset dataset for MIST top model variations trained. All models are trained without any auxiliary MAGMa loss and without any simulated spectra added to the training set. "MIST - unfolding (linear)" indicates a model without unfolding layers featuring only a single linear layer to predict fingerprint outputs. "MIST - unfolding (FFN)" indicates a MIST model trained without unfolding layers, but instead using a dense feedforward network of 3 layers. Each experiment was conducted on a single split of the data. Results are shown ± the standard error of the mean.

Model	Cosine similarity	Log likelihood
MIST	$\mathbf{0.6818 \pm 0.0066}$	$\mathbf{-0.0248 \pm 0.0005}$
MIST - unfolding (linear)	0.6701 ± 0.0061	-0.0250 ± 0.0005
MIST - unfolding (FFN)	0.6617 ± 0.0063	-0.0255 ± 0.0005

S1.5 MAGMa substructure labels

The MAGMa algorithm is used to take pairs of (molecule, spectrum) from the trained dataset and attribute molecule substructures to MS2 peaks [6]. In

brief, the algorithm works by progressively removing atoms from the original structure to create substructures. Each time an atom is removed, the resulting subfragments are given a tolerance of $\pm H$ to account for hydrogen rearrangements. We highlight several spectra from the public GNPS dataset to illustrate this substructure labeling (Fig. S7b-d).

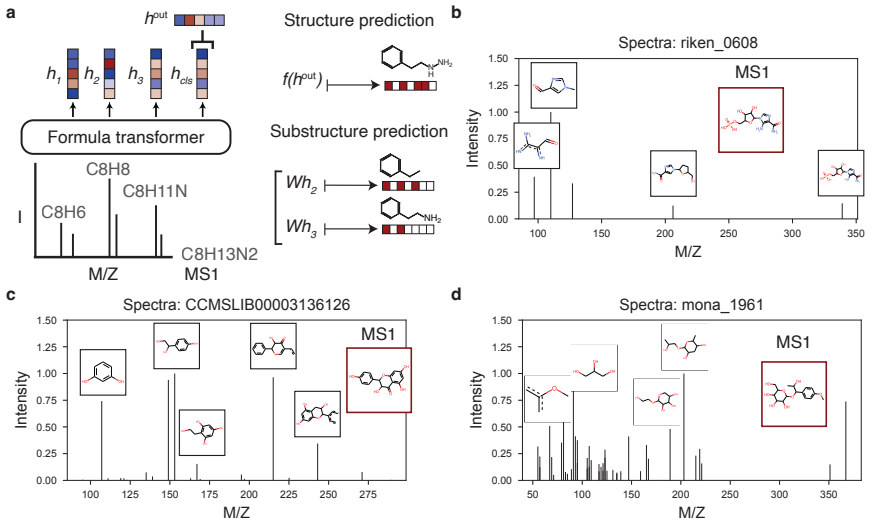

Fig. S7 Labeling data substructures. a. MIST is trained to build representations of each peak in the spectrum. The model uses the precursor peak representation to predict a fingerprint of the full molecule, and when substructures can be labeled for other peaks, the model learns to predict these as well. **b-d.** Example substructure peak labels in the training set computed with MAGMa. The full compound is outlined in red. Other substructures for the top 6 highest intensity peaks are outlined in black.

S1.6 Alternative transformer baselines

Contemporaneously with this work, ref. [7] have proposed to utilize a Transformer architecture applied to m/z values. In their architecture, m/z values are first transformed into frequency encodings using a sinusoidal embedding, inspired by the original positional encodings of the Transformer architecture [8]. Because the code is not available at the time of this work, we develop our

own in-house implementation of this approach and hyperparameter optimize model parameters on our public-domain dataset, settling upon a 256 hidden size, 4 transformer layers, 0.3 dropout, 0 weight decay, and 0.0002 learning rate.

We find that in our data regime, the multi-scale Transformer performs equivalently to the FFN baseline, as shown in Table S5. We hypothesize that this is due to the relatively small amount of data in our training set compared to the original authors and also the lack of explicit information about neutral loss relationships between peaks.

S1.7 Spectra noising

To increase generality as in MetFID [1], we introduce an algorithm to probabilistically add noise to input training spectra (Alg. S1). Spectra are selected for augmentation with a 50% probability. For each spectrum selected for augmentation, we define two augmentation operations: `remove`, in which a peak is fully removed, and `intensity`, in which a peak's intensity is multiplied by a randomly sampled scalar value $s_i \sim \mathcal{N}(1, 1)$, truncated to a minimum of 0. The probability of removing each peak, p_j is proportional to its intensity value to ensure high intensity peaks are less likely to be removed:

$$p_j(\text{remove}) = 1 - p_j(\text{keep}), \qquad p_j(\text{keep}) \propto \exp[I_j] \tag{30}$$

For each noised spectrum, we sample the number of peaks to remove from a binomial distribution with $p_{\text{remove}} = 0.5$ and we sample a second binomial distribution for the number of peaks to rescale with probability $p_{\text{intensity}} = 0.1$.

Algorithm S1 Noising a training spectrum

procedure NOISESPECTRUM($\mathcal{S}$)
  $\mathcal{S}$ # Example spectrum
  $I \leftarrow$ `get_intens`($\mathcal{S}$) # Get all intensities
  peaks $\leftarrow |\mathcal{S}|$
  if `rnd()` ≥ 0.5 **then**
    num_remove $\leftarrow$ `Binomial`(peaks, 0.5)
    num_rescale $\leftarrow$ `Binomial`(peaks, 0.1)
    for j in $\mathcal{S}$ **do**
      modify_probs$_j \leftarrow$ `exp`(I_j)
    end for
    remove_inds $\leftarrow$ `Sample`(peaks, num_remove, modify_probs)
    rescale_inds $\leftarrow$ `Sample`(peaks,num_rescale, modify_probs)
    $\mathcal{S} \leftarrow$ `remove_peaks`($\mathcal{S}$, remove_inds)
    $\mathcal{S} \leftarrow$ `rescale_peaks`($\mathcal{S}$, rescale_inds)
  end if
end procedure

S1.8 Hyperparameter optimization

The tree-structured Parzen estimator hyperparameter search algorithm [9] as implemented by Optuna [10] was used to identify the best set of hyperparameters for each model. Hyperparameters were optimized once to maximize performance on a single validation split of the data, excluding auxilary loss terms. RayTune [11] was used to enable efficient parallelization across 3 RTX 3090 GPUs and a HyperBand scheduler was used to prune unpromising hyperparameters for efficiency. We ensured the same compute resources were available for tuning our models and each of the baselines and therefore capped resources at 100 trials on 3 GPUs. We conducted two separate hyperparameter sweeps to identify model parameters best for predicting fingerprints from CSI:FingerID (NIST dataset) and also the Morgan 4096 bit fingerprints (CANOPUS benchmark dataset). Parameters are described in Table S11 and Table S12.

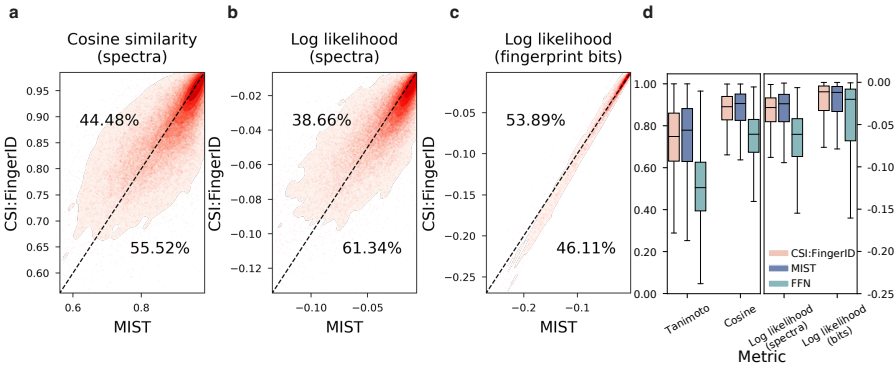

Fig. S8 Single MIST model comparison to CSI:FingerID. Molecular fingerprints are predicted by MIST and CSI:FingerID for every spectrum in the test set using a single MIST model as in Figure 2. The performance for each spectrum by cosine similarity to the true fingerprint **(a)** or log likelihood **(b)** is evaluated and plotted. Points below the line represent instances where MIST is more performant. **c.** Equivalent evaluation showing the likelihood of predicting each fingerprint bit correctly across all spectra. **d.** The performance of CSI:FingerID, MIST, and FFN, a baseline inspired by MetFID, are shown. "Tanimoto," "Cosine," and "Log likelihood (spectra)" indicate performance across spectra and "Log likelihood (bits)" indicates performance across bits (higher is better). Median lines are shown; boxes show the interquartile range, with whiskers indicating 1.5x interquartile ranges. Statistics are pooled across $n = 18,700$ test spectra (tanimoto, cosine, and spectra log likelihood) and $n = 5,496$ fingerprint bits (bit log likelihood). All results are aggregated across 3 splits of the data.

Table S3 Full performance metrics for CSI:FingerID, FFN, MIST, and a MIST ensemble of 5 models are shown averaged across 3 structural splits of the data. Log likelihoods are clamped such that they have a minimum of −5. Results are shown ± standard error with an arbitrary number of significant figures.

	Tanimoto similarity (spectra)	Cosine similarity (spectra)	Log likelihood (spectra)	Log likelihood (bits)
CSI:FingerID	0.7382 ± 0.0011	0.8759 ± 0.0006	-0.0352 ± 0.0002	-0.0352 ± 0.0008
FFN	0.5108 ± 0.0012	0.7418 ± 0.0009	-0.0710 ± 0.0003	-0.0710 ± 0.0016
MIST	0.7475 ± 0.0012	0.8774 ± 0.0007	-0.0345 ± 0.0002	-0.0345 ± 0.0007
MIST (5x)	$\mathbf{0.7629 \pm 0.0012}$	$\mathbf{0.8892 \pm 0.0006}$	$\mathbf{-0.0309 \pm 0.0002}$	$\mathbf{-0.0309 \pm 0.0007}$

Table S4 Full retrieval accuracy performance metrics for CSI:FingerID, FFN, MIST, and a MIST ensemble of 5 models are shown averaged across 3 structural splits of the data.

	Top 1(%)	Top 5(%)	Top 10(%)	Top 20(%)	Top 50(%)	Top 100(%)	Top 200(%)
CSI:FingerID + Cosine	38.24	68.88	77.07	83.20	89.12	92.23	94.76
CSI:FingerID + Bayes	**43.00**	**74.77**	**82.15**	87.23	91.70	94.13	96.00
MIST + Cosine	29.72	59.56	69.26	73.60	84.10	88.13	91.84
MIST + Contrastive	34.45	69.28	79.04	85.75	91.63	94.56	96.49
MIST (5x) + Cosine	31.70	62.85	72.33	79.05	85.92	89.76	93.00
MIST (5x) + Contrastive	37.39	72.90	82.03	**88.20**	**93.32**	**95.75**	**97.40**

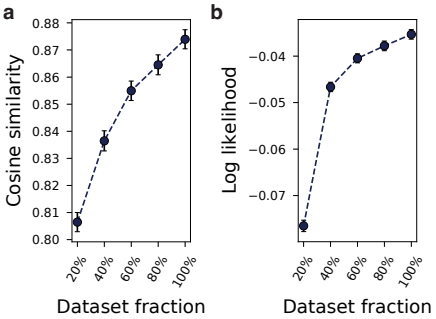

Fig. S9 Dataset ablations on the CSI2022 dataset validate that model performance is limited by data set size. MIST fingerprint prediction accuracy improves as a function of dataset size in terms of both (**a**) cosine similarity and (**b**) log likelihood. Model performance is computed on the partially proprietary CSI2022 dataset. Dataset ablations are conducted for a single split of the data. Log likelihood values are clamped to a minimum of −5. Error bars show 95% confidence intervals for the mean (n=3,116 test spectra). For this ablation, MIST models are trained without simulated data and without MAGMa auxiliary loss to reduce model training time.

Table S5 Full fingerprint prediction accuracy on CANOPUS GNPS subset dataset for various model ablations. All model performances are computed on a merged dataset of fingerprint predictions computed for 3 independent random structural split hold outs of the CANOPUS dataset. FFN: Feed forward neural network. Transformer: A partial reimplementation of [7]. Results are shown plus or minus the standard error of the mean.

Model	Cosine similarity	Log likelihood
FFN	0.5368 ± 0.0042	-0.0344 ± 0.0004
Transformer [7]	0.5267 ± 0.0039	-0.0333 ± 0.0003
MIST - unfolding	0.6796 ± 0.0035	-0.0244 ± 0.0003
MIST - simulated	0.6880 ± 0.0038	-0.0248 ± 0.0003
MIST - MAGMa	0.6907 ± 0.0038	$\mathbf{-0.0228 \pm 0.0003}$
MIST - pairwise	$\mathbf{0.6963 \pm 0.0037}$	-0.0232 ± 0.0003
MIST	0.6953 ± 0.0037	-0.0236 ± 0.0003

Table S6 Fingerprint prediction accuracy on primary CSI2022 dataset to compare models with and without pairwise featurizations. Model performance values are computed on a merged dataset of fingerprint predictions computed for 3 independent random structural split test sets. All MIST models described were trained without simulated data or MAGMa auxiliary loss features for quicker model runs. Log likelihoods are clamped such that they have a minimum of -5. Results are shown ± standard error with an arbitrary number of significant figures.

Model	Cosine similarity	Log likelihood
MIST + pairwise	0.8693 ± 0.0007	-0.0368 ± 0.0002
MIST - pairwise	0.8688 ± 0.0008	-0.0369 ± 0.0002

Table S7 Fingerprint prediction accuracy on CANOPUS GNPS subset dataset for MIST models with various featurizations of peak intensities. All models are trained without any auxiliary MAGMa loss and without any simulated spectra added to the training set. "MIST + float intensity" indicates a concatenation of a single floating point value of intensity between 0 and 1; "MIST + float intensity & pooling" indicates a concatenation of a floating point intensity and selective intensity-weighted pooling of hidden states in the transformer; "MIST + log intensity" indicates a concatenation of a single log-transformed intensity value; "MIST + cat. intensity" indicates a concatenation of a one-hot vector of intensity transformed into one of 10 evenly spaced categorical bins; "MIST + zero intensity" indicates a concatenation of a constant float of zero rather than the true intensity. Each experiment was conducted on a single split of the data and average results are shown ± the standard error of the mean.

Model	Cosine similarity	Log likelihood
MIST + float intensity	$\mathbf{0.6818 \pm 0.0066}$	$\mathbf{-0.0248 \pm 0.0005}$
MIST + float intensity & pooling	0.6775 ± 0.0063	-0.0262 ± 0.0005
MIST + log intensity	0.6743 ± 0.0065	-0.0264 ± 0.0005
MIST + cat. intensity	0.6743 ± 0.0066	-0.0252 ± 0.0005
MIST + zero intensity	0.6726 ± 0.0063	-0.02707 ± 0.0005

Table S8 Full retrieval accuracy on CANOPUS GNPS subset dataset for various model ablations. All model performance values are computed on a merged dataset of retrieval predictions computed for 3 independent random structural splits of 3 random structural splits of the CANOPUS dataset. FFN: Feed forward neural network.

	Top 1(%)	Top 5(%)	Top 10(%)	Top 20(%)	Top 50(%)	Top 100(%)	Top 200(%)
FFN fingerprint	17.309	37.121	45.611	54.634	63.946	71.329	77.359
FFN contrastive	20.632	44.996	54.389	63.536	75.062	80.558	86.095
MIST contrastive - pretrain	24.774	50.656	60.624	68.827	78.384	83.593	88.146
MIST contrastive	28.384	55.373	65.217	72.970	81.255	85.480	89.377
MIST fingerprint	29.368	55.332	63.536	72.231	80.476	85.726	89.418
MIST contrastive + fingerprint	**30.703**	**58.120**	**68.927**	**75.709**	**84.094**	**87.916**	**92.355**

Table S9 Average molecular similarity between spectra pairs with the 0.1% highest similarity according to the specified distance function.

Method	Average Tanimoto Similarity
Upper bound	0.451
MIST	0.324
MS2DeepScore [12]	0.219
Spec2Vec [13]	0.186
Modified Cosine [14]	0.133
Random	0.076

Table S10 Forward model ablations demonstrate the utility of unfolding and reverse mass predictions. Three variants of the forward simulation model are trained on a single split of the public GNPS dataset: "FFN" (full model), "FFN - unfolding" (full model a single linear layer to map to the 15,000 dimension output), and "FFN - reverse" (full model without predicting mass differences as in [15]). The average cosine similarity and coverage (i.e., fraction of overlapping peaks between the top 100 predicted peaks and ground truth spectra) between the binned spectra and true spectra are shown. Values are plotted ± the standard error of the mean.

Model	Cosine similarity	Coverage
FFN	0.3839 ± 0.0056	0.5815 ± 0.0086
FFN - unfolding	0.3502 ± 0.0058	0.5555 ± 0.0088
FFN - reverse	0.2583 ± 0.0058	0.5160 ± 0.0088

Table S11 Model hyperparameters searched and set for CSI fingerprint prediction.

Model	**Parameter**	**Grid**	**Value**
Forward	learning rate	$[1e-4, 1e-3]$	0.00086
	scheduler	[True, False]	False
	learning rate decay	$[0.7, 0.999]$	-
	dropout	$\{0.1, 0.2, 0.3\}$	0.2
	hidden size, d	$\{64, 128, 256, 512\}$	512
	layers, l	$\{1, 2, 3\}$	2
	unfolding	[True, False]	True
	unfolding layers, γ	$[1, 2, 3, 4]$	3
	unfolding weight, λ_{fwd}	$[1e-4, 1]$	0.003
	batch size	-	64
	total parameters	-	88.8M
FFN	learning rate	$[1e-5, 1e-3]$	0.00087
	scheduler	[True, False]	False
	learning rate decay	$[0.7, 0.999]$	-
	dropout	$\{0.1, 0.2, 0.3\}$	0.0
	hidden size, d	$\{64, 128, 256, 512\}$	512
	layers, l	$\{1, 2, 3\}$	2
	spectra bins	$[1000, 2000, \ldots 15000]$	11000
	batch size	-	64
	total parameters	-	8.7M
MIST	learning rate	$[1e-5, 1e-3]$	0.00078
	scheduler	[True, False]	False
	weight decay	$\{1e-6, 1e-7, 0\}$	$1e-7$
	learning rate decay	$[0.7, 0.999]$	-
	dropout	$\{0.1, 0.2, 0.3\}$	0.1
	hidden size, d	$\{64, 128, 256, 512\}$	256
	layers, l	$\{1, 2, 3, 4, 5\}$	2
	fraction real data f_{real}	$\{0.1, 0.2, \ldots, 1.0\}$	0.6
	unfolding layers, ν	$\{1, 2, 3, 4, 5\}$	4
	unfolding loss weight, λ_{unfold}	$\{0.1, 0.2, \ldots, 1.0\}$	0.4
	magma loss weight, λ_{magma}	$\{1, 2, \ldots, 15\}$	8
	probability noised spectrum, p_{noise}	-	0.5
	probability remove peak, p_{remove}	-	0.5
	probability rescale peak, $p_{\text{intensity}}$	-	0.1
	batch size	-	128
	total parameters	-	15.2M
Contrastive	learning rate	$[1e-5, 1e-3]$	0.00022
	scheduler	[True, False]	True
	weight decay	$\{1e-6, 1e-7, 0\}$	0
	learning rate decay	$[0.7, 0.999]$	0.89
	contrastive weight, λ_c	$\{0.1, 0.2, \ldots, 1.0\}$	0.2
	fraction real data f_{real}	$\{0.1, 0.2, \ldots, 1.0\}$	0.2
	probability noised spectrum, p_{noise}	-	0.5
	probability remove peak, p_{remove}	-	0.5
	probability rescale peak, $p_{\text{intensity}}$	-	0.1
	batch size	-	32
	total parameters	-	16.2M

Table S12 Model hyperparameters searched and set for Morgan fingerprint prediction.

Model	**Parameter**	**Grid**	**Value**
Forward	learning rate	$[1e-4, 1e-3]$	0.00054
	scheduler	[True, False]	False
	learning rate decay	$[0.7, 0.999]$	-
	dropout	$\{0.1, 0.2, 0.3\}$	0.3
	hidden size, d	$\{64, 128, 256, 512\}$	512
	layers, l	$\{1, 2, 3\}$	2
	unfolding	[True, False]	True
	unfolding layers, γ	$[1, 2, 3, 4]$	3
	unfolding weight, λ_{fwd}	$[1e-4, 1]$	0.0016
	batch size	-	64
	total parameters	-	83.7M
FFN	learning rate	$[1e-5, 1e-3]$	0.00087
	scheduler	[True, False]	False
	learning rate decay	$[0.7, 0.999]$	-
	dropout	$\{0.1, 0.2, 0.3\}$	0.3
	hidden size, d	$\{64, 128, 256, 512\}$	512
	layers, l	$\{1, 2, 3\}$	2
	spectra bins	$[1000, 2000, \ldots 15000]$	11000
	batch size	-	64
	total parameters	-	8.0M
MIST	learning rate	$[1e-5, 1e-3]$	0.0003
	scheduler	[True, False]	False
	weight decay	$\{1e-6, 1e-7, 0\}$	$1e-7$
	learning rate decay	$[0.7, 0.999]$	-
	dropout	$\{0.1, 0.2, 0.3\}$	0.0
	hidden size, d	$\{64, 128, 256, 512\}$	512
	layers, l	$\{1, 2, 3, 4, 5\}$	3
	fraction real data f_{real}	$\{0.1, 0.2, \ldots, 1.0\}$	0.1
	unfolding layers, ν	$\{1, 2, 3, 4, 5\}$	5
	unfolding loss weight, λ_{unfold}	$\{0.1, 0.2, \ldots, 1.0\}$	0.1
	magma loss weight, λ_{magma}	$\{1, 2, \ldots, 15\}$	4
	probability noised spectrum, p_{noise}	-	0.5
	probability remove peak, p_{remove}	-	0.5
	probability rescale peak, $p_{\text{intensity}}$	-	0.1
	batch size	-	128
	total parameters	-	25.6M
Contrastive	learning rate	$[1e-5, 1e-3]$	0.0002
	scheduler	[True, False]	True
	weight decay	$\{1e-6, 1e-7, 0\}$	$1e-7$
	learning rate decay (10k steps)	$[0.7, 0.999]$	0.89
	contrastive weight, λ_c	$\{0.1, 0.2, \ldots, 1.0\}$	0.2
	fraction real data f_{real}	$\{0.1, 0.2, \ldots, 1.0\}$	0.2
	probability noised spectrum, p_{noise}	-	0.5
	probability remove peak, p_{remove}	-	0.5
	probability rescale peak, $p_{\text{intensity}}$	-	0.1
	batch size	-	32
	total parameters	-	27.7M

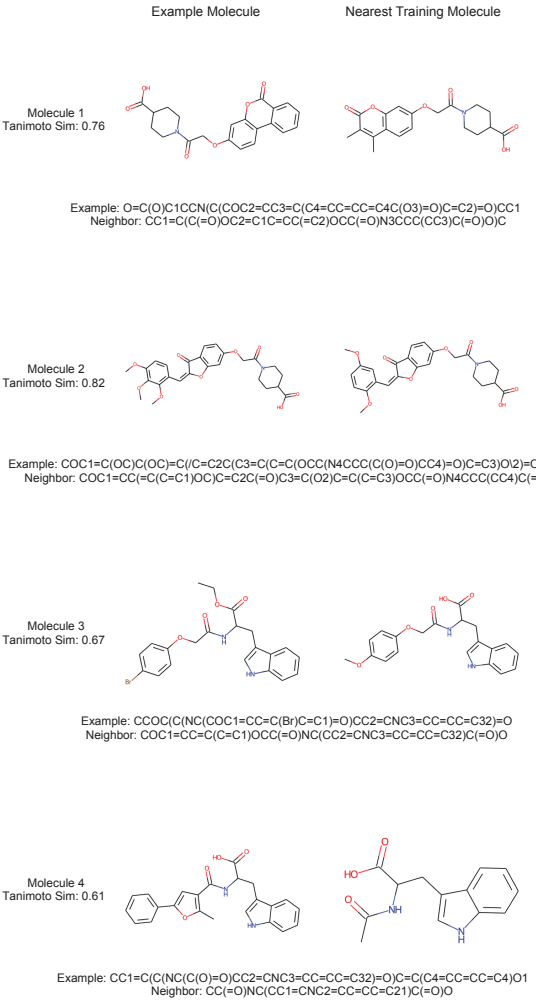

Fig. S10 Example molecules and their training precedents for Fig. 3. Each molecule is shown along side its nearest neighbor in the training set, with its Tanimoto similarity listed.

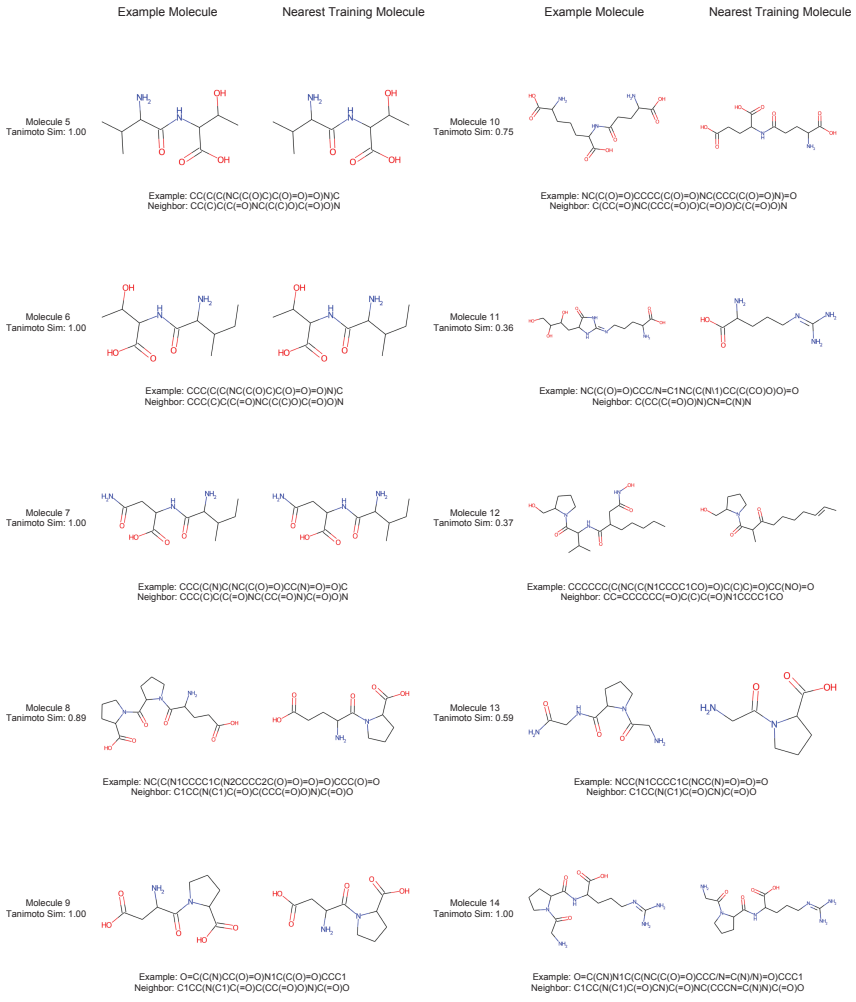

Fig. S11 Example molecules and their training precedents for Fig. 4. Each molecule is shown along side its nearest neighbor in the training set, with its Tanimoto similarity listed. Even in cases for which the proposed molecular structure appears in the training set, it is predicted on the basis of a new experimental spectrum that does not exactly match the reference spectrum.

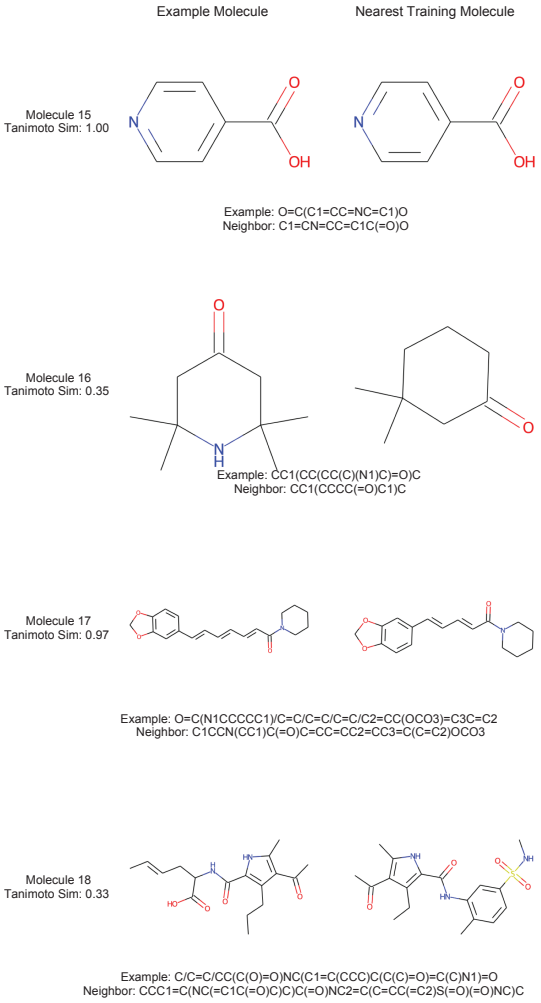

Fig. S12 Example molecules and their training precedents for Fig. 5. Each molecule is shown along side its nearest neighbor in the training set, with its Tanimoto similarity listed. Even in cases for which the proposed molecular structure appears in the training set, it is predicted on the basis of a new experimental spectrum that does not exactly match the reference spectrum.

References

[1] Fan, Z., Alley, A., Ghaffari, K. & Ressom, H. W. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics* **16**, 104 (2020). URL https://doi.org/10.1007/s11306-020-01726-7.

[2] Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **16**, 299–302 (2019). URL https://www.nature.com/articles/s41592-019-0344-8.

[3] Wishart, D. S. *et al.* HMDB: the human metabolome database. *Nucleic Acids Research* **35**, D521–D526 (2007). ISBN: 1362-4962 Publisher: Oxford University Press.

[4] Kim, H. W. *et al.* NPClassifier: a deep neural network-based structural classification tool for natural products. *Journal of Natural Products* **84**, 2795–2807 (2021). ISBN: 0163-3864 Publisher: ACS Publications.

[5] Mills, R. H. *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome link Bacteroides vulgatus proteases with disease severity. *Nature Microbiology* **7**, 262–276 (2022). ISBN: 2058-5276 Publisher: Nature Publishing Group.

[6] Ridder, L., van der Hooft, J. J. & Verhoeven, S. Automatic compound annotation from mass spectrometry data using MAGMa. *Mass Spectrometry* **3**, S0033–S0033 (2014). ISBN: 2187-137X Publisher: The Mass Spectrometry Society of Japan.

[7] Voronov, G. *et al.* Multi-scale sinusoidal embeddings enable learning on high resolution mass spectrometry data (2022). URL https://arxiv.org/abs/2207.02980.

[8] Vaswani, A. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* **30** (2017).

[9] Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems* **24** (2011).

[10] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631 (2019).

[11] Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).

[12] Huber, F., van der Burg, S., van der Hooft, J. J. & Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics* **13**, 1–14 (2021). ISBN: 1758-2946 Publisher: Springer.

[13] Huber, F. *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology* **17**, e1008724 (2021). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008724. Publisher: Public Library of Science.

[14] Huber, F. *et al.* matchms-processing and similarity evaluation of mass spectrometry data. *bioRxiv* (2020). Publisher: Cold Spring Harbor Laboratory.

[15] Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science* **5**, 700–708 (2019). ISBN: 2374-7943 Publisher: ACS Publications.