Emergence of molecular structures from repository-scale self-supervised learning on tandem mass spectra

Roman Bushuiev[1,2†], Anton Bushuiev[2†], Raman Samusevich[1,2], Corinna Brungs[1], Josef Sivic[2*] and Tomáš Pluskal[1*]

[1]Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences.
[2]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University.

*Corresponding author(s). E-mail(s): josef.sivic@cvut.cz; tomas.pluskal@uochb.cas.cz;
Contributing authors: roman.bushuiev@uochb.cas.cz; anton.bushuiev@cvut.cz;
†These authors contributed equally to this work.

Abstract

Tandem mass spectrometry (MS/MS) is the primary method for characterizing biological and environmental samples at a molecular level. Despite this, the interpretation of tandem mass spectra remains a challenge. Existing computational methods for predictions from mass spectra heavily rely on limited spectral libraries and on hard-coded human expertise. Here we introduce a transformer-based neural network pre-trained in a self-supervised way on millions of unannotated tandem mass spectra from our new GeMS (GNPS Experimental Mass Spectra) dataset mined from the MassIVE GNPS repository. We show that pre-training our model to predict masked spectral peaks and chromatographic retention orders leads to the emergence of rich representations of molecular structures, which we name DreaMS (Deep Representations Empowering the Annotation of Mass Spectra). Fine-tuning the pre-trained neural network to predict spectral similarity, molecular fingerprints, chemical properties, and the presence of fluorine from tandem mass spectra yields state-of-the-art performance across all the tasks. This underscores the practical utility of DreaMS across diverse mass spectrum interpretation tasks and establishes it as a stepping stone for future advances in the field. We make our new dataset and pre-trained models available to the community and release the DreaMS Atlas – a molecular network of 201 million MS/MS spectra constructed using DreaMS annotations.

Keywords: Mass spectrometry, metabolomics, machine learning, self-supervised learning, large language models

Introduction

The discovery and identification of small molecules and metabolites have a profound impact on various scientific fields, including drug development [1], environmental analysis [2], and disease diagnosis [3]. However, only a tiny fraction of natural small molecules have been discovered to date, estimated to be less than 10% of those present in the human body or the entire plant kingdom [4]. The vast majority of the natural chemical space thus remains unexplored.

Tandem mass spectrometry coupled with liquid chromatography (LC–MS/MS) is a central analytical technique for investigating the molecular composition of biological and environmental samples. When analyzing a sample, the LC–MS/MS system separates molecules through liquid chromatography, ionizes them, and records their mass-to-charge ratios (m/z), generating a series of mass spectra (referred to as MS^1). Each MS^1 spectrum is acquired at a specific retention time (RT) and represents the abundance of ions in terms of their m/z ratios (i.e., peaks). Using a technique referred to as data-dependent acquisition (DDA), selected ions (referred to as precursor ions) undergo fragmentation (typically using collision-induced dissociation, CID), yielding additional tandem mass spectra (referred to as MS^2 or MS/MS), where signals characterize molecular fragments of a single selected ion. Although MS^2 and deeper MS^n tandem mass spectra constitute the primary source of structural information in mass spectrometry, their interpretation remains exceptionally challenging. In particular, a mere 2% of MS/MS spectra can be annotated with molecular structures using reference spectral libraries [5, 6], and less than 10% of MS/MS spectra can typically be annotated using state-of-the-art machine learning tools [7].

Existing methods for the interpretation of mass spectra can be classified into three major categories: spectral similarity, forward annotation, and inverse annotation. Spectral similarity algorithms aim to define a similarity measure on mass spectra, which reflects the similarity of the underlying molecular structures. Classic dot-product-based algorithms are optimized for querying spectral libraries and linking spectra of similar compounds into molecular networks [8–10]. Unsupervised shallow machine learning methods, MS2LDA [11] and spec2vec [12], aim to devise more versatile spectral similarities based on statistical occurrences of spectral peaks. By contrast, recently developed contrastive learning approaches aim to explicitly approximate similarities in molecular structures [13–15]. The utility of similarity-based methods is heavily dependent on the richness of annotated spectral libraries, which are, however, inherently limited in size [16]. Therefore, forward annotation methods seek to extend MS/MS datasets with *in silico* spectra by simulating CID fragmentation of molecules via combinatorial optimization based on hand-crafted priors [17, 18] or graph neural networks [19–21]. Contrastingly, inverse annotation methods aim to directly annotate spectra with molecular structures,

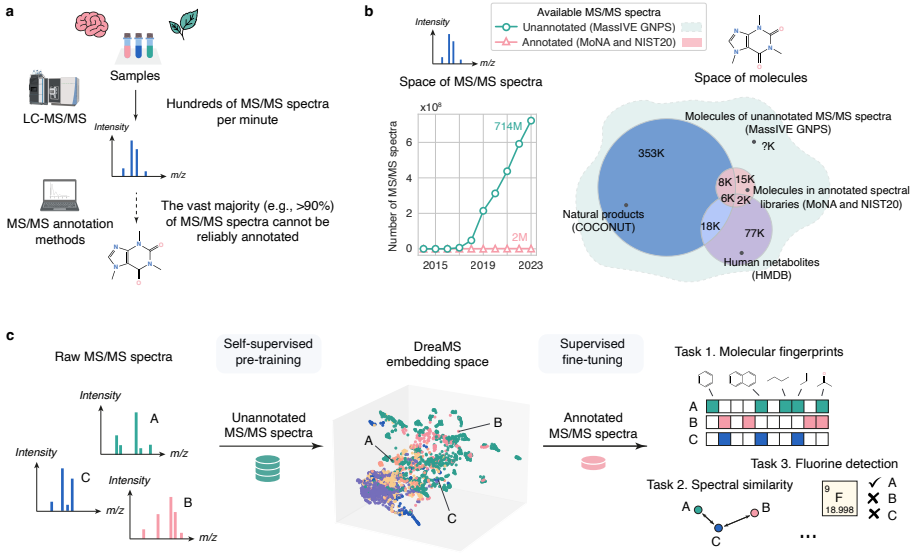


Fig. 1 The DreaMS neural network overcomes the limitation of mass spectral libraries. a, Given a biological or environmental sample, the LC–MS/MS system produces hundreds of mass spectra (MS/MS) per minute, characterizing its molecular composition. However, less than 10% of MS/MS spectra can typically be assigned with molecular structures using existing annotation methods. **b,** Even though the number of publicly available unannotated experimental mass spectra has been rapidly growing over recent years (left; green), annotated spectral libraries are still highly limited in terms of both the number of spectra (left; pink) and the coverage of molecular structures (right; Venn diagram). State-of-the-art annotation methods rely on spectral libraries as training or retrieval datasets. By contrast, we base our method on training from vast unannotated MS/MS datasets, assuming that the molecular coverage of these data surpasses spectral libraries (right; dashed green shape). **c,** We propose the DreaMS neural network, which is capable of learning molecular representations from raw unannotated mass spectra through self-supervised learning. After being pre-trained in a self-supervised way, DreaMS can be fine-tuned for a wide range of spectrum annotation problems via supervised transfer learning, leveraging spectral libraries as well as other sources of annotated data.

either in the approximate form of molecular fingerprints [22], molecular formulas [23, 24], chemical properties [25, 26], or as complete *de novo* molecular structures [27, 28].

The most prominent and well-established method for the interpretation of mass spectra, SIRIUS [29], comprises a pipeline of approximate inverse annotation tools based on combinatorics, discrete optimization, and machine learning leveraging mass spectrometry domain expertise. First, it explains a given MS/MS spectrum with a fragmentation tree by assigning chemical formulas to individual spectral peaks [24]. Then, it employs a series of support vector machines (SVMs) with kernels, designed to operate on mass spectra and fragmentation trees. SIRIUS predicts a proprietary CSI:FingerID fingerprint [22], which is used to retrieve a molecular structure from a compound database such as PubChem [30]. Recently developed competitive methods,

MIST [31] and MIST-CF [32], replace crucial components of SIRIUS with neural networks trained on spectral libraries. Both methods employ a similar transformer architecture which operates on chemical formulas assigned to individual peaks as input tokens. Whereas MIST-CF assigns chemical formulas via energy-based modeling, MIST uses these formulas to predict a molecular fingerprint and employs it to retrieve a molecular structure from compound databases. To achieve a level of performance that is competitive with SIRIUS, both methods employ additional domain-specific computationally demanding components such as mass decomposition [33], data pseudo-annotation with the forward annotator MAGMA [34], or the generation of *in silico* spectral libraries [31]. The reliance of the state-of-the-art machine learning models on a variety of auxiliary methods suggests that the capacity of training spectral libraries is the principal bottleneck of the process. In fact, the molecular structures of the standard training spectral libraries MoNA [35] and NIST20 [36] cover only a limited subset of known natural molecules (Fig. 1b), not to mention the vastness of the chemical space that remains to be explored.

In this study, we introduce a large self-supervised neural network (with 116 million parameters) trained directly on the repository-scale collection of raw experimental mass spectra (Fig. 1b,c). Inspired by the remarkable achievements of large transformer models pre-trained on biological protein sequences [37–41], text [42, 43], and images [44], we developed a transformer model for tandem mass spectrometry named DreaMS (Deep Representations Empowering the Annotation of Mass Spectra). Without relying on prior methodologies or human domain expertise, DreaMS can be easily adapted to a wide range of spectrum annotation tasks and thus act as a foundation model for tandem mass spectrometry [45]. To achieve this, we first constructed a high-quality dataset, GeMS (GNPS Experimental Mass Spectra), comprising up to 700 million MS/MS spectra mined from the GNPS repository [46]. Second, we designed a transformer neural network and pre-trained it on our GeMS data to predict masked spectral peaks and chromatographic retention orders. We show that through optimization towards these self-supervised objectives on unannotated mass spectra, our model discovers rich representations of molecular structures. Specifically, we find that the DreaMS representations are organized according to the structural similarity between molecules and are robust to mass spectrometry conditions. Finally, we demonstrate that DreaMS, fine-tuned for diverse mass spectrum annotation tasks, including the prediction of spectral similarity, molecular fingerprints, chemical properties, and the presence of fluorine, surpasses both traditional algorithms and recently developed machine learning models.

Results

New datasets of MS/MS spectra for deep learning

Comprehensive and high-quality datasets are essential for effective self-supervised learning [47–49]. However, spectral libraries of metabolites are

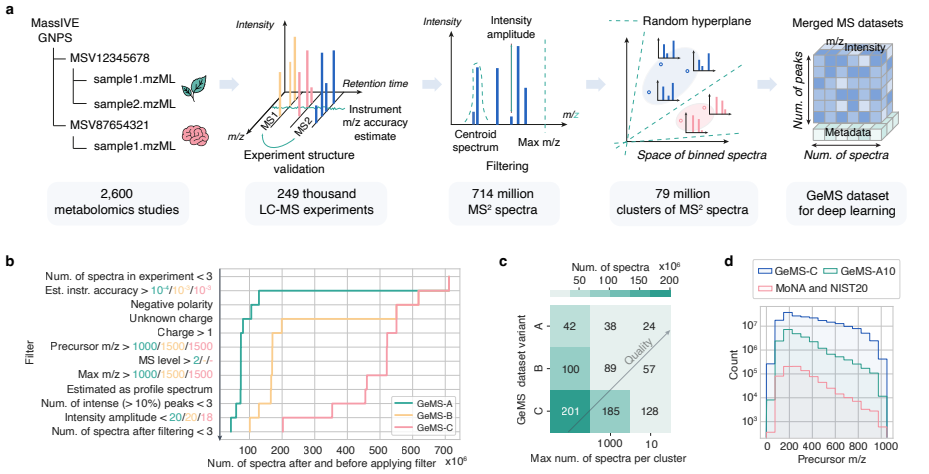


Fig. 2 GeMS – high-quality datasets of unannotated MS/MS spectra from GNPS. a, The workflow of mining GeMS datasets from the GNPS repository. MS/MS spectra from metabolomics studies were filtered using experiment- and spectrum-level quality criteria, clustered with locality-sensitive hashing, and packed into a tensor-like dataset suitable for deep learning. **b,** Quality criteria defining the A, B, and C subsets of GeMS data, ordered from top to bottom by the sequence of their application. **c,** Sizes of the nine final clustered and unclustered GeMS variants. Each cell in the heatmap corresponds to a specific variant, denoted in the text as, for instance, GeMS-A10, based on the respective axes. 79 million clusters on top represent the fully clustered GeMS-C1 subset. **d,** All the GeMS dataset variants are orders of magnitude larger than the union of MoNA and NIST20 spectral libraries and cover a wide range of molecular masses.

limited in size and only cover a tiny fraction of the entire chemical space. To the best of our knowledge, there are no large standardized datasets of mass spectra suitable for unsupervised or self-supervised deep learning. Therefore, we mine the MassIVE GNPS repository [46] to establish a new large-scale and high-quality dataset comprising hundreds of millions of experimental MS/MS spectra, which we name GeMS (GNPS Experimental Mass Spectra).

Our mining pipeline consists of five main steps (Fig. 2a): First, we collected 250 thousand LC–MS/MS experiments from diverse biological and environmental studies, covering virtually the entire GNPS part of the MassIVE repository [46]. Second, we extracted from these experiments approximately 700 million MS/MS spectra. Next, we developed a pipeline of quality control algorithms allowing us to filter the collected spectra into three subsets: GeMS-A, GeMS-B, and GeMS-C, each with consecutively larger size at the expense of quality. The quality criteria include, for example, the estimation of the instrument m/z accuracy associated with a single LC–MS/MS experiment or the number of high-intensity signals within each spectrum (Fig. 2b). Subsequently, we addressed redundancy in GeMS by clustering similar spectra using locality-sensitive hashing (LSH). The LSH algorithm approximates cosine similarity, a common metric for identifying similar spectra, but operates in linear time, enabling efficient clustering of our large-scale data. Specifically, we restricted

the number of spectra per cluster to a certain quantity, such as 10 or 1,000, resulting in a total of nine different GeMS dataset variants (Fig. 2c). Finally, we stored the GeMS spectra, including selected LC–MS/MS metadata, in our compact HDF5-based binary format designed for deep learning (Extended Data Table 4). Our new GeMS datasets are orders of magnitude larger (Fig. 2d) than existing spectral libraries and are well organized in a tensor-shaped structure, unlocking new possibilities for repository-scale metabolomics research [50, 51]. The details of the data collection and filtering are provided in Online Methods.

Self-supervised pre-training on tandem mass spectra

Leveraging the GeMS-A10 dataset, our highest-quality subset of GeMS, we propose DreaMS – a self-supervised model which learns molecular representations directly from unannotated mass spectra. Self-supervision is a form of unsupervised learning, where the training objective typically involves a reconstruction of corrupted data points. This approach has been demonstrated to yield rich representations (i.e., embeddings) of words, images, or proteins, which effectively generalize across diverse tasks [38, 42, 44]. However, self-supervised learning on mass spectra of small molecules has not been explored yet, primarily because of the absence of large standardized datasets and strong inductive biases for large-scale learning. We have addressed this challenge by designing a transformer-based neural network for MS/MS spectra and training it using our new large-scale data.

The core of our self-supervised approach (Figure 3a) is BERT-style [42] spectrum-to-spectrum masked modeling. We represent each spectrum as a set of two-dimensional continuous tokens associated with pairs of peak m/z and intensity values. Then we mask a fraction (30%) of random m/z ratios from each set (or spectrum), sampled proportionally to corresponding intensities, and train the model to reconstruct each masked peak. Additionally, we introduce an extra token, which we refer to as the precursor token. This token is never masked and contains a precursor ion m/z ratio and a precursor-specific artificial intensity value, serving as an aggregator of spectrum-level information into a single embedding, akin to a sentence-level token or a graph-level master node in the related language of graph models [42, 52]. Besides masked m/z prediction, we employ a retention order training objective. Each training example is formed as a pair of partially masked spectra, sampled from the same LC–MS/MS experiment, and the neural network simultaneously learns to reconstruct the masked peaks and to predict which one elutes first in chromatography.

The backbone of our DreaMS neural network architecture (Figure 3b) is based on the transformer encoder [53]. It consists of a sequence of multi-head self-attention blocks, which gradually derive the representations of peaks and relationships between them. We adjust the standard architecture to handle high-resolution molecular masses. First, each m/z value is preprocessed with a modification of Fourier features, a computer vision technique shown to improve

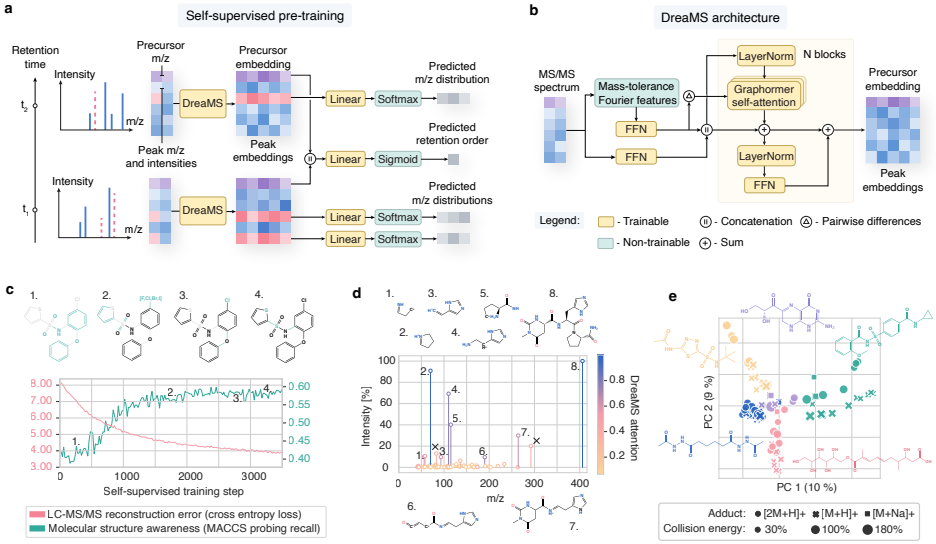


Fig. 3 The DreaMS neural network discovers molecular structures through self-supervised learning on mass spectra. a, Self-supervision setup. The DreaMS neural network is provided with a pair of spectra (blue) from the same LC–MS/MS experiment along with their precursor m/z values (purple). A portion of m/z ratios in both spectra is masked (red), and the model is trained to reconstruct these values by predicting a probability distribution over m/z ratios for each mask. Additionally, the model learns to predict the retention order of the two spectra (i.e., the probability that $t_2 > t_1$). **b,** Architecture of the DreaMS neural network. Initially, input spectral peaks, including precursor m/z with artificial intensity value, are assigned mass-tolerance Fourier features and processed with shallow feed-forward neural networks (FFN). The subsequent transformer encoder, equipped with Graphormer self-attention layers operating on pair-wise mass differences, refines the encoded peaks into high-dimensional output embeddings. **c,** Emergence of molecular structures from self-supervised training. At each self-supervised training step, DreaMS parameters are frozen, and a separate linear layer is trained to predict interpretable MACCS keys fingerprints from precursor peak embeddings, allowing the inspection of learned molecular fragments. As the self-supervised loss decreases (red), the recall in MACCS bits increases (green), indicating the model's ongoing discovery of new molecular structures. The MACCS fragments for individual bits are visually presented on top. **d,** An example spectrum colored based on the maximum attention value across all attention heads for each peak (blue indicates high attention, yellow indicates low attention). DreaMS learns to focus on high-intensity peaks representing fragments and to ignore noise. Molecules depict fragment annotations produced by Mass Frontier (Thermo Fisher Scientific); crossed intense peaks lack annotations. **e,** Principal component analysis (PCA) applied to selected precursor embeddings demonstrates the linear clustering of mass spectra according to molecular structures, remaining robust to multiple ionization adducts and normalized collision energies (NCE) associated with each molecule.

the representation of high-resolution details in images [54]. In essence, each m/z value is decomposed into pre-defined sine and cosine frequencies capturing both the integer and the floating-point part of a single mass. We additionally process the Fourier features with a feed-forward network to enable, for example, the learning of possible elemental compositions associated with input masses [55]. We incorporate intensity values by processing them through a shallow

feed-forward network and then concatenating them with the processed Fourier features. This combined representation serves as the input for the transformer. Second, we explicitly feed differences in Fourier features between all pairs of peaks to self-attention heads, following the Graphormer architecture [56]. This enables the transformer to attend directly to neutral losses without increasing computational complexity through the introduction of extra tokens or modifications to the dot-product attention mechanism. Finally, instead of treating masked m/z prediction as a regression problem, we treat it as classification and train the model to predict a probability distribution over a binned mass range for each mask. This approach allows the network to model the uncertainty of predictions when multiple m/z values could match the same intensity.

We hypothesize that when the DreaMS model is trained to predict masked m/z ratios and chromatographic retention orders, it implicitly learns to reason in terms of molecular structures. To test this hypothesis empirically, we first employed a machine learning technique called linear probing [57] to assess the evolution of learned representations during training. Specifically, when training a simple linear regression from precursor embeddings to interpretable MACCS keys fingerprints [58] at each training step, we noted that during self-supervised training, the model progressively discovers molecular fragments (Fig. 3c). Second, our analysis of transformer attention heads revealed that the model learned to prioritize peaks representing molecular structures and to ignore noise (Fig. 3d). Third, we found that the DreaMS representation space linearly clustered spectra according to molecular structures, even when fragmented under different ionization and fragmentation settings (Fig. 3e). Ablation studies of the DreaMS components indicate that pre-training on the high-quality GeMS-A10 dataset, mass-tolerance Fourier features, and a masked m/z objective formulated as classification rather than regression are key components of our self-supervised approach (Extended Data Fig. 4).

Transfer learning to MS/MS spectrum annotation tasks

The emergence of molecular structures in DreaMS is a result of self-supervised training from extensive unannotated mass spectral data, without relying on annotated MS/MS libraries, chemical databases or human expertise. It motivates us to investigate DreaMS as a foundation model possessing a general understanding of molecules, which can be transferred to various spectrum annotation tasks. In particular, we have adapted the network to the prediction of spectral similarity, molecular fingerprints, chemical properties, and the identification of fluorine-containing molecules. For each task, we augment the pre-trained model with a simple linear head and fine-tune the entire neural network end to end on annotated spectral libraries. To ensure the generalization of fine-tuned models beyond spectral libraries, we halt fine-tuning when the model's performance plateaus on validation spectra of molecules with different Murcko histograms from those in the training set (except for the fingerprint prediction benchmark established by Goldman et al. [31]). A Murcko histogram is our new molecular representation, generalizing the notion

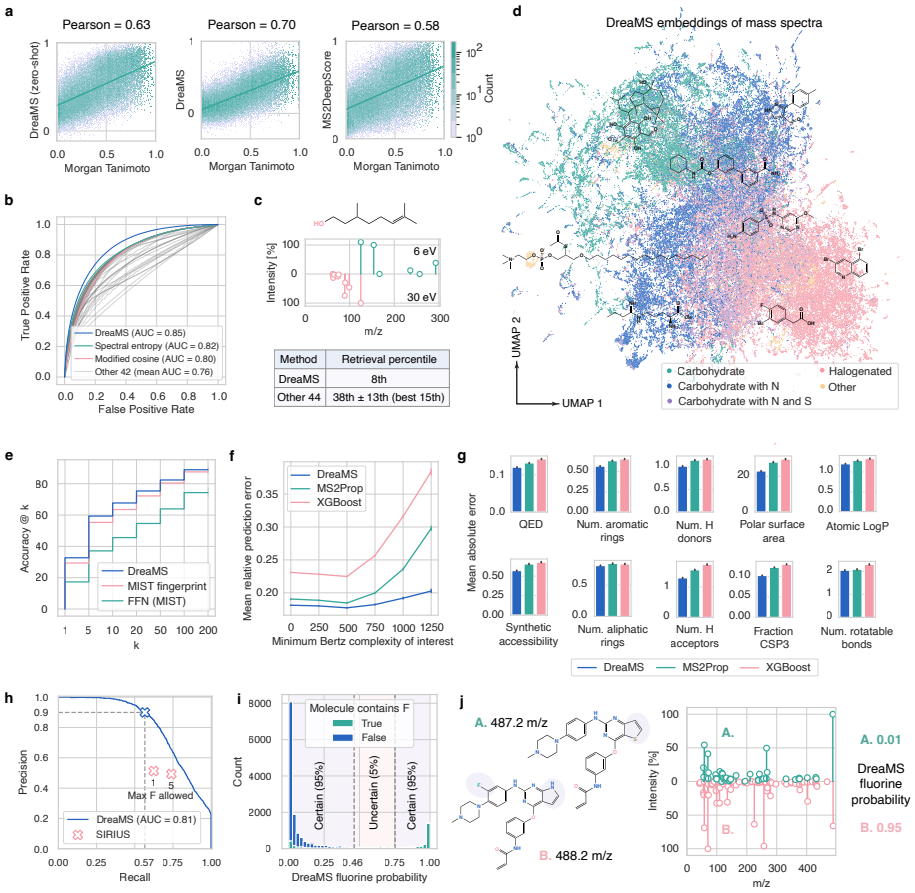


Fig. 4 The DreaMS neural network outperforms state-of-the-art methods at solving a variety of spectrum annotation tasks. a, Zero-shot (i.e., unsupervised) cosine similarity of DreaMS representations outperforms MS2DeepScore [13] in predicting precursor Tanimoto similarities. Contrastive fine-tuning further enhances the correlation (fine-tuned models are referred to simply as DreaMS; Extended Data Fig. 3a). **b,** Fine-tuned cosine similarity outperforms standard spectral similarity algorithms in a library retrieval task [9]. **c,** For a selected query spectrum (green) and a distinct candidate spectrum of the same molecule (pink), DreaMS retrieves the candidate at a low 8th percentile (in the distribution of all evaluation pairs) whereas all classic methods fail to recognize these spectra as representing the same compound (Extended Data Fig. 3c). **d,** UMAP projection of DreaMS embeddings reveals the organization of representation space according to molecular formulas. **e,** DreaMS fine-tuned to predict Morgan fingerprints (blue) outperforms the MIST fingerprint model [31] (pink), as well as the feed-forward neural network baseline (green), in terms of compound database retrieval accuracy on the MIST benchmark [31]. **f–g,** DreaMS outperforms existing chemical property prediction models [25, 26] (error bars show 99% confidence intervals of 1,000 bootstraps). Notably, by predicting Bertz complexity, DreaMS excels on practically interesting, high-complexity examples. **h,** DreaMS (blue) surpasses SIRIUS (pink; two different settings) in detecting fluorinated molecules, achieving almost two-fold greater 90% precision under 57% recall. **i,** Choosing another threshold for 90% precision in fluorine absence predictions categorizes predictions as certain (95% of DreaMS predictions) or uncertain. **j,** Model generalization demonstrated on two similar spectra of nearly identical molecules with different fluorine annotations. DreaMS confidently predicts correct annotations despite the absence of similar training examples (Extended Data Fig. 3d). The details on the evaluation datasets and metrics are provided in Online Methods.

of a Murcko scaffold [59] (described in Online Methods and Extended Data Fig. 1). This universal transfer learning protocol consistently yields models with state-of-the-art performance across different tasks, eliminating the need for constructing task-specific components or extensively tuning model hyperparameters (Fig. 4).

The first task we tackle is spectral similarity, which can be performed directly in the space of DreaMS representations. Remarkably, we observe that even before any fine-tuning, cosine similarity in the embedding space outperforms the cutting-edge supervised algorithm MS2DeepScore [13] in terms of correlation with molecular similarity measures (Fig. 4a). This result emphasizes the amount of information captured by self-supervised representations, especially when considering the fact that MS2DeepScore was explicitly trained on pairs of annotated spectra to approximate their corresponding molecular similarities. Nevertheless, we find that simple zero-shot similarity of DreaMS often lacks sensitivity to small differences in molecular structures (Extended Data Fig. 3b), which are typically crucial for spectral library retrieval and molecular networking. To address this limitation, we disentangle the embeddings of similar molecules through a short but accurate contrastive fine-tuning on hard examples. These examples consist of triplets comprising a reference spectrum, a different positive spectrum of the same molecular structure, and a negative spectrum of a molecule with a different structure but a similar mass, differing by no more than 0.05 Da from a reference molecule. During fine-tuning, the model refines DreaMS representations by bringing the reference-positive pairs closer together than the reference-negative pairs. We use only a subset of 5,500 molecules from MoNA to avoid biasing the DreaMS representations towards spectral libraries. In a challenging scenario of retrieving similar or different molecules within the 10-ppm precursor m/z difference, fine-tuned DreaMS significantly outperforms 44 standard spectral similarity measures (Figure 4b). The contrastive fine-tuning procedure not only increases sensitivity to details but also globally enhances the correlation with molecular similarities, despite not being explicitly optimized for it (Figure 4a). The analysis of the resultant embeddings with UMAP projections [60] reveals that the DreaMS representations are organized by chemical formulas and structural motifs of the underlying molecules (Fig. 4d, Extended Data Fig. 2). Notably, we find that averaging DreaMS embeddings across samples yields embeddings capturing the composition of complete metabolic profiles (Fig. 5). To the best of our knowledge, there are no existing tools that enable the direct comparison of metabolomes corresponding to different samples or species.

The second problem we address is predicting Morgan fingerprints from mass spectra and using them to retrieve molecules from PubChem. Importantly, in contrast to prior work, our method is capable of predicting fingerprints directly from raw spectra in a single forward pass. This breaks the dependency of machine learning on computationally-heavy intermediate steps such as the assignment of chemical formulae to individual input peaks or the combinatorial generation of fragmentation trees. We find that the fine-tuned DreaMS neural

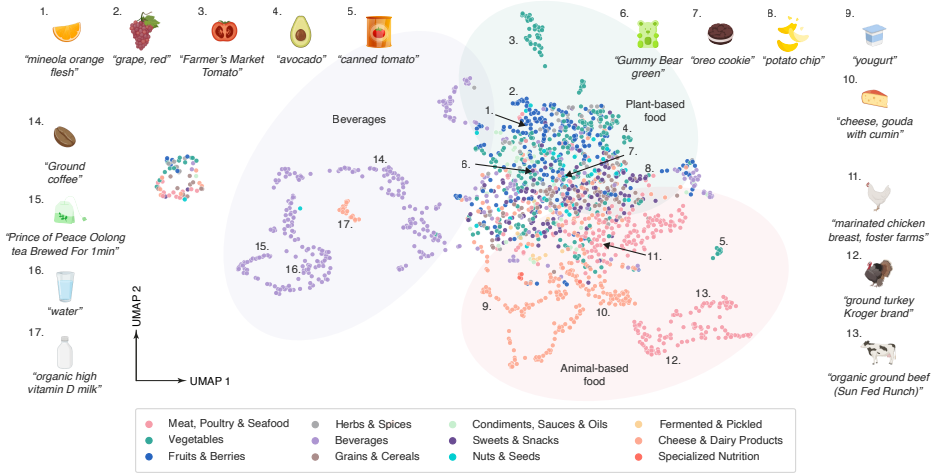


Fig. 5 Sample-average DreaMS embeddings enable the sample-level analysis of metabolomics data, as exemplified on food LC–MS/MS datasets. Each point on the UMAP plot represents a centroid of DreaMS embeddings (i.e., mean embedding values across dimensions) of all MS/MS spectra acquired from a certain food sample [61]. Numbered points indicate selected example samples and refer to their textual descriptions assigned by the data collectors. The figure demonstrates that the space of sample-level embeddings correctly captures the taxonomy of food items presented to DreaMS as collections of MS/MS spectra. Specifically, the space is organized into three major regions predominantly populated with beverages (purple ellipse), plant food items (green ellipse), and animal food items (pink ellipse). Beverages are separated into milk beverages (orange) and other beverages (purple). Animal-based food items are divided into clusters comprising various dairy products (orange) and types of meat (pink). Plant-based food items show less distinction between categories and are primarily classified as vegetables (green), fruits (blue), and herbs and spices (grey). Individual categories (colors) were assigned to sample descriptions using ChatGPT 4 [62].

network outperforms the state-of-the-art deep learning model MIST in the retrieval of molecular structures using predicted fingerprints (Fig. 4e, Extended Data Table 1), despite the latter is based on molecular formulae assigned to individual spectral peaks.

The third problem we tackle is predicting molecular properties of practical interest. Specifically, fast and precise prediction of pharmaceutically relevant chemical properties, such as those involved in Lipinski's rule of five [63], is essential for the large-scale screening of drug candidates [25]. Similarly, the prediction of Bertz molecular complexity from mass spectra is a promising way to search for biosignatures beyond Earth [26]. The rich molecular knowledge encoded in DreaMS and its fast inference time inspire us to explore the direct prediction of these properties, bypassing the determination of complete molecular structures. We fine-tune the DreaMS neural network to simultaneously predict these and a series of other molecular characteristics. Our model achieves state-of-the-art performance on the prediction of all properties considered for fine-tuning (Fig. 4f-g).

Finally, we address the task of detecting fluorinated molecules from their mass spectra. Currently, there is no practically applicable method capable of detecting fluorine with high precision [64]. This task is particularly challenging because fluorine has only one stable isotope and because fluorinated ions do not exhibit well-defined fragmentation patterns. The state-of-the-art method SIRIUS relies on combinatorial search of fragmentation rules, resulting in a high number of false-positive predictions and requiring extensive runtime. To overcome this limitation, we fine-tune DreaMS to predict the probability of fluorine presence. We evaluate our method on 17,000 previously unreported MS/MS spectra from our in-house library. Whereas SIRIUS does not exceed a precision value of 0.51, DreaMS achieves a precision of 0.91 with a recall of 0.57 and surpasses SIRIUS in recall at low precision values (Fig. 4h). This high precision without a significant drop in recall on a large test dataset ensures the practical applicability of our method, suggesting that fluorine detections by DreaMS are predominantly correct, and the model confidently identifies every second fluorinated molecule (Fig. 4i). We additionally demonstrate the strong generalization capacity of our fine-tuned model by identifying correct and confident detection of fluorine for spectra of molecules structurally distinct from all training examples (Fig. 4j).

DreaMS Atlas – repository-scale molecular network

Large-scale metabolomics research is currently constrained by the processing time of spectrum annotation methods. Consequently, the only methods that are practically applicable on a large scale are variations of MASST [50, 67], a traditional modified cosine similarity search algorithm optimized for quickly identifying nearly identical spectra. By contrast, our fully neural network-based models for interpreting MS/MS spectra are both computationally efficient and versatile, enabling the annotation of approximately one million spectra per hour on a GPU machine (8x NVIDIA A100). Therefore, we utilize our fine-tuned models to annotate 201 million mass spectra from the MassIVE GNPS repository (covering virtually all positive-mode metabolomics spectra) with DreaMS predictions and organize them into a comprehensive molecular network, which we name the DreaMS Atlas (Fig. 6a).

The DreaMS Atlas is constructed as an approximate five-nearest-neighbor graph based on GeMS mass spectra. Each node represents a DreaMS embedding of a mass spectrum, while each edge represents a DreaMS similarity between the corresponding nodes. To enhance the representativeness and reduce redundancy, we compute the graph for a subset of 34 million mass spectra, which represents 201 million spectra in GeMS-C, clustered based on LSH hashes and DreaMS similarities (details are provided in Online Methods). We populate each node with DreaMS molecular property and fluorine presence predictions, as well as MassIVE metadata such as the study descriptions and species information. When constructing the graph, we additionally include nodes corresponding to the embeddings of mass spectra from the MoNA and NIST20 spectral libraries.

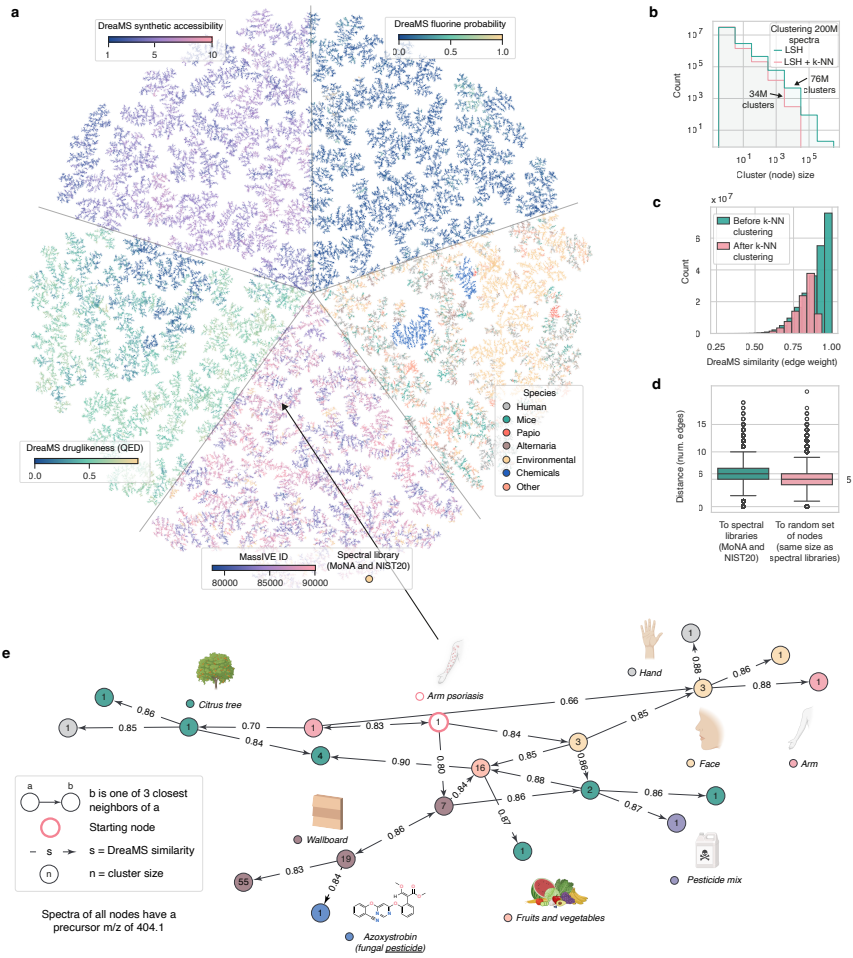


Fig. 6 The DreaMS Atlas, a molecular network of 201 million MS/MS spectra, offers a comprehensive systematization of the entire MassIVE GNPS repository. The DreaMS Atlas is built as a five-nearest-neighbor NN-Descent graph [65] (5-NN) from DreaMS embedding similarities between MS/MS spectra from GeMS-C1, MoNA, and NIST20. Each node includes DreaMS-based property predictions (e.g., druglikeness) and MassIVE metadata (e.g., species). **a,** TMAP projection [66] of the 5-NN graph, divided into five pieces, showcasing the different types of node annotations. A subset of 1 million GeMS-C1 nodes and all MoNA and NIST20 nodes are shown. **b,** Each node represents a cluster of mass spectra. First, GeMS-C1 spectra are GeMS-C LSH cluster representatives (green). Second, neighborhoods with DreaMS similarity >0.9 were collapsed to single nodes (pink). The distribution of cluster sizes follows the inverse polynomial trend as the depicted log-log histograms exhibit a linear trend. **c,** The DreaMS Atlas is predominantly populated with high-similarity edges, indicating effective interpolation between spectra of different molecules via transitive connections (green, pink). Neighborhood clustering effectively eliminated the vast majority of nearly-identical spectra (pink). **d,** Spectral libraries are distributed relatively evenly across the DreaMS Atlas, as shown by the median distance of six edges to a random set of nodes. Combining observations from **b, c, d**, the DreaMS Atlas systematizes the dark metabolome of MassIVE GNPS enabling the interpretation of spectra via short high-similarity paths. **e,** Directed three-hop neighborhood of a selected node illustrates such interpretation and highlights the DreaMS Atlas as a research hypothesis generator connecting distinct scientific studies. Specifically, a spectrum from the arm psoriasis study links to the spectrum of the fungicide azoxystrobin from MoNA, suggesting a potential link between psoriasis and the fungicide, abundantly found in various other environmental and biological samples.

Moving on to the analysis of the global composition of the DreaMS Atlas, approximately 33% of nodes represent clusters with more than one spectrum. These clusters presumably correspond to individual molecules, with the largest cluster comprising 393 thousand spectra from 23 thousand distinct LC–MS/MS experiments. The distribution of cluster sizes follows an inverse polynomial trend (Fig. 6b). Regarding edges, the network exhibits strong connectivity, with the majority (67%) of edges displaying high similarities (> 0.8), as depicted in Fig. 6c. Simultaneously, 99.7% of nodes form a single connected component of the graph, despite a total of sixteen thousand components. These findings suggest that the DreaMS Atlas enables effective interpolation between spectra of different molecules through strongly connected transitive paths between nodes, even when considering only the five closest neighbors.

This observation motivates us to investigate the connectivity between arbitrary GeMS spectra and spectral library entries. Despite the limited size of the libraries, we find that they are distributed relatively evenly across the DreaMS Atlas. The median distance from a randomly sampled node to any MoNA or NIST20 spectrum is six edges, compared to the median distance of five edges to a random subset of nodes of the same size (Fig. 6d). This observation aligns with the previous analysis of spectral library composition in terms of molecules, in comparison with natural product structures [68]. Such a distribution of annotated nodes with respect to DreaMS similarities suggests that many spectra from MassIVE GNPS can be interpreted by propagating spectral library annotations [69] or interpolating between them. On the other hand, nodes distant from spectral libraries or arbitrarily sampled nodes (upper box plot outliers in Fig. 6d) may represent structurally novel molecules [51]. Ultimately, the DreaMS Atlas can function as a database that can be efficiently queried or populated with new spectra.

We demonstrate the interpretation of a mass spectrum by propagating through its neighbors in the DreaMS Atlas. In particular, we consider a spectrum from an arm psoriasis LC–MS/MS study. Autoimmune diseases such as psoriasis are characterized by complex etiology, which remains incompletely understood [70]. We illustrate how the diversity of the DreaMS Atlas facilitates the exploration of these factors by connecting various scientific studies. Specifically, our analysis reveals a potential association between psoriasis and the fungicide azoxystrobin (Fig. 6e), which, to the best of our knowledge, has not been previously reported. The DreaMS Atlas neighbors also suggest that exposure to azoxystrobin may occur through various environmental sources such as contaminated food, treated trees or mold and mildew-resistant wallboards, thereby supporting recent hypotheses regarding the origin of the fungicide in samples from children and pregnant women [71].

Discussion

In this article we introduce DreaMS, a universal transformer model for interpreting tandem mass spectra. First, we show that through self-supervised

pre-training on GeMS, our new large collection of unannotated MS/MS spectra from the GNPS part of MassIVE, the DreaMS neural network acquires embeddings of mass spectra that reflect underlying molecular structures. Second, we demonstrate the effective fine-tuning capability of DreaMS for a diverse range of mass spectrum annotation problems, achieving state-of-the-art performance across all evaluated tasks. Finally, we present DreaMS Atlas – a comprehensive molecular network constructed using DreaMS annotations for 201 million mass spectra from GeMS.

Although our results strongly indicate the emergence of molecular structure knowledge from training on raw, unannotated mass spectra, the full potential of this approach remains to be unlocked. In particular, we trained our model using only a subset of available mass spectra. Scaling the self-supervised learning to larger datasets (e.g., by mining spectra from additional repositories such as MetaboLights [72]) and incorporating more diverse mass spectrometry data (e.g., including spectra beyond positive ionization modes or singly charged precursor ions) is expected to yield even richer representations of mass spectra, potentially even more accurately capturing the structures of underlying molecules. Additionally, our method is focused solely on tandem mass spectra, disregarding other important features such as MS^1 isotopic patterns or adduct distributions, which are important, for example, for correct chemical formula determination [23].

Our work opens up new possibilities in two directions of metabolomics-related research. First, we have introduced a general data-driven transformer model that can be tailored to virtually any mass spectrum interpretation task, thereby moving away from traditional hand-crafted or rule-based approaches for individual problems. Now that we have made our pre-trained model available to the community, we anticipate that it will serve as a foundational tool, providing a starting point (i.e., a base model or a feature extractor) for developing more powerful neural network architectures. Second, we have introduced the DreaMS Atlas, a comprehensive resource enabling the interpretation of mass spectra by leveraging DreaMS predictions and MassIVE GNPS metadata for 201 million mass spectra. Treating the DreaMS Atlas as an approximation of the space of chemically plausible molecular structures offers new perspectives on various challenges of computational chemistry. For example, fragment-based drug design could be addressed by interpolating between known drugs in the DreaMS Atlas, while the detection of novel structurally unique compounds with potentially original modes of action can be facilitated by identifying sparsely connected regions in the graph structure of the DreaMS Atlas. Ultimately, annotation of the DreaMS Atlas using a DreaMS model successfully fine-tuned for *de novo* structure generation has the potential to significantly expand our knowledge and understanding of the still largely unexplored chemical space.

Data Availability

The GeMS datasets, DreaMS Atlas, and weights of pre-trained models can be accessed through our GitHub repository (https://github.com/pluskal-lab/DreaMS). Our in-house data for fluorine detection evaluation is available under the MassIVE accession number MSV000094528 (https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=676a38e2dd574a15905e807d78cf1e57), and the food datasets are available at MSV00008490 (https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=ce3254fe529d43f48077d7ad55b7da09). The MoNA spectral library can be downloaded from the official website (https://mona.fiehnlab.ucdavis.edu/), while the NIST20 library is not publicly available due to licensing restrictions.

Code Availability

The source code for data preparation, model training, and experiments is available at our GitHub repository (https://github.com/pluskal-lab/DreaMS).

Acknowledgments

We thank Serena Khoo, Peter G Mikhael, Regina Barzilay, Kai Dürkop, Sebastian Böcker, Juho Rousu, Robin Schmid, Téo Hebra, Tito Damiani, Joshua Smith, Ratthachat Chatpatanasiri, Niek de Jonge for fruitful discussions on our work. We thank Samuel Goldman for helping to reproduce the MIST evaluation benchmark. We thank Fred Rooks for editing this manuscript. We also thank the Dagstuhl (especially Dagstuhl Seminar #22181) and Shonan (especially Shonan Seminar #179) computational metabolomics communities for insightful discussions and brainstormings on various aspects of mass spectrometry and machine learning. Icons in figures were created using BioRenderer (www.biorender.com) and Flaticon (www.flaticon.com).

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140). T.P. is supported by the Czech Science Foundation (GA CR) grant 21-11563M and by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 891397. C.B. was supported by the Czech Academy of Sciences PPLZ fellowship number L200552251. This work was also co-funded by the European Union (ERC, FRONTIER, 101097822). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Conflict of interest statement

T.P. is a co-founder of mzio GmbH, which develops technologies related to mass spectrometry data processing.

Author contributions

R.B., A.B., R.S., J.S., T.P. designed the research. R.B., A.B. implemented the methods and performed the computational analyses. R.B., A.B., J.S., T.P. wrote the manuscript. C.B. prepared a new experimental dataset for benchmarking fluorine detection.

References

[1] Atanasov, A. G. *et al.* Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* **20** (3), 200–216 (2021). URL https://doi.org/10.1038/s41573-020-00114-z. https://doi.org/10.1038/s41573-020-00114-z .

[2] Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367** (6476), 392–396 (2020). URL https://www.science.org/doi/abs/10.1126/science.aay3164. https://doi.org/10.1126/science.aay3164, https://arxiv.org/abs/https://www.science.org/doi/pdf/10.1126/science.aay3164 .

[3] Banerjee, S. Empowering clinical diagnostics with mass spectrometry. *ACS Omega* **5** (5), 2041–2048 (2020). URL https://doi.org/10.1021/acsomega.9b03764. https://doi.org/10.1021/acsomega.9b03764 .

[4] Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature Methods* **18** (7), 747–756 (2021). URL https://doi.org/10.1038/s41592-021-01197-1. https://doi.org/10.1038/s41592-021-01197-1 .

[5] da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* **112** (41), 12549–12550 (2015). URL https://www.pnas.org/doi/abs/10.1073/pnas.1516878112. https://doi.org/10.1073/pnas.1516878112, https://arxiv.org/abs/https://www.pnas.org/doi/pdf/10.1073/pnas.1516878112 .

[6] Vinaixa, M. *et al.* Mass spectral databases for lc/ms- and gc/ms-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry* **78**, 23–35 (2016). URL https://www.sciencedirect.com/science/article/pii/S0165993615300832. https://doi.org/https://doi.org/10.1016/j.trac.2015.09.005 .

[7] de Jonge, N. F. *et al.* Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* **18** (12), 103 (2022). URL https://doi.org/10.1007/s11306-022-01963-y. https://doi.org/10.1007/s11306-022-01963-y .

[8] Bittremieux, W. *et al.* Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. *Journal of the American Society for Mass Spectrometry* **33** (9), 1733–1744 (2022). URL https://doi.org/10.1021/jasms.2c00153. https://doi.org/10.1021/jasms.2c00153 .

[9] Li, Y. *et al.* Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods* **18** (12), 1524–1531 (2021). URL https://doi.org/10.1038/s41592-021-01331-z. https://doi.org/10.1038/s41592-021-01331-z .

[10] Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using gnps. *Nature Protocols* **15** (6), 1954–1991 (2020). URL https://doi.org/10.1038/s41596-020-0317-5. https://doi.org/10.1038/s41596-020-0317-5 .

[11] van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences* **113** (48), 13738–13743 (2016). URL https://www.pnas.org/doi/abs/10.1073/pnas.1608041113. https://doi.org/10.1073/pnas.1608041113, https://arxiv.org/abs/https://www.pnas.org/doi/pdf/10.1073/pnas.1608041113 .

[12] Huber, F. *et al.* Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology* **17** (2), 1–18 (2021). URL https://doi.org/10.1371/journal.pcbi.1008724. https://doi.org/10.1371/journal.pcbi.1008724 .

[13] Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics* **13** (1), 84 (2021). URL https://doi.org/10.1186/s13321-021-00558-4. https://doi.org/10.1186/s13321-021-00558-4 .

[14] Voronov, G. *et al.* Multi-scale sinusoidal embeddings enable learning on high resolution mass spectrometry data. *CoRR* **abs/2207.02980** (2022). URL https://doi.org/10.48550/arXiv.2207.02980. https://doi.org/10.48550/ARXIV.2207.02980, https://arxiv.org/abs/2207.02980 .

[15] Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. *Nature Methods* **19** (6), 675–678 (2022). URL https://doi.org/10.1038/s41592-022-01496-1. https://doi.org/10.1038/s41592-022-01496-1 .

[16] Bittremieux, W., Wang, M. & Dorrestein, P. C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics* **18** (12), 94 (2022). URL https://doi.org/10.1007/

s11306-022-01947-y. https://doi.org/10.1007/s11306-022-01947-y .

[17] Wang, F. *et al.* Cfm-id 4.0: More accurate esi-ms/ms spectral prediction and compound identification. *Analytical Chemistry* **93** (34), 11692–11700 (2021). URL https://doi.org/10.1021/acs.analchem.1c01465. https://doi.org/10.1021/acs.analchem.1c01465 .

[18] Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* **8** (1), 3 (2016). URL https://doi.org/10.1186/s13321-016-0115-9. https://doi.org/10.1186/s13321-016-0115-9 .

[19] Murphy, M. *et al.* Krause, A. *et al.* (eds) *Efficiently predicting high resolution mass spectra with graph neural networks.* (eds Krause, A. *et al.*) *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Vol. 202 of *Proceedings of Machine Learning Research*, 25549–25562 (PMLR, 2023). URL https://proceedings.mlr.press/v202/murphy23a.html.

[20] Goldman, S., Li, J. & Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks (2023). 2304.13136.

[21] Goldman, S., Bradshaw, J., Xin, J. & Coley, C. W. Prefix-tree decoding for predicting mass spectra from molecules (2023). 2303.06470.

[22] Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences* **112** (41), 12580–12585 (2015). URL https://www.pnas.org/doi/abs/10.1073/pnas.1509788112. https://doi.org/10.1073/pnas.1509788112, https://arxiv.org/abs/https://www.pnas.org/doi/pdf/10.1073/pnas.1509788112 .

[23] Xing, S., Shen, S., Xu, B., Li, X. & Huan, T. Buddy: molecular formula discovery via bottom-up ms/ms interrogation. *Nature Methods* **20** (6), 881–890 (2023). URL https://doi.org/10.1038/s41592-023-01850-x. https://doi.org/10.1038/s41592-023-01850-x .

[24] Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *Journal of Cheminformatics* **8** (1), 5 (2016). URL https://doi.org/10.1186/s13321-016-0116-8. https://doi.org/10.1186/s13321-016-0116-8 .

[25] Voronov, G. *et al.* Ms2prop: A machine learning model that directly predicts chemical properties from mass spectrometry data for novel compounds. *bioRxiv* (2022). URL https://www.biorxiv.org/content/early/2022/10/11/2022.10.09.511482. https://doi.org/10.1101/2022.10.09.511482, https://arxiv.org/abs/https://www.biorxiv.org/content/early/2022/10/11/2022.10.09.511482.full.pdf .

[26] Gebhard, T. D. *et al.* Inferring molecular complexity from mass spectrometry data using machine learning (2022) .

[27] Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. Msnovelist: de novo structure generation from mass spectra. *Nature Methods* **19** (7), 865–870 (2022). URL https://doi.org/10.1038/s41592-022-01486-3. https://doi.org/10.1038/s41592-022-01486-3 .

[28] Butler, T. *et al.* Ms2mol: A transformer model for illuminating dark chemical space from mass spectra (2023). https://doi.org/10.26434/chemrxiv-2023-vsmpx-v2 .

[29] Dührkop, K. *et al.* Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **16** (4), 299–302 (2019). URL https://doi.org/10.1038/s41592-019-0344-8. https://doi.org/10.1038/s41592-019-0344-8 .

[30] Hoffmann, M. A. *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology* **40** (3), 411–421 (2022). URL https://doi.org/10.1038/s41587-021-01045-9. https://doi.org/10.1038/s41587-021-01045-9 .

[31] Goldman, S. *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence* **5** (9), 965–979 (2023). URL https://doi.org/10.1038/s42256-023-00708-3. https://doi.org/10.1038/s42256-023-00708-3 .

[32] Goldman, S., Xin, J., Provenzano, J. & Coley, C. W. Mist-cf: Chemical formula inference from tandem mass spectra. *Journal of Chemical Information and Modeling* (2023). URL https://doi.org/10.1021/acs.jcim.3c01082. https://doi.org/10.1021/acs.jcim.3c01082 .

[33] Dührkop, K., Ludwig, M., Meusel, M. & Böcker, S. Darling, A. & Stoye, J. (eds) *Faster mass decomposition.* (eds Darling, A. & Stoye, J.) *Algorithms in Bioinformatics*, 45–58 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

[34] Ridder, L. *et al.* Substructure-based annotation of high-resolution multistage msn spectral trees. *Rapid Communications in Mass Spectrometry* **26** (20), 2461–2471 (2012). URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/rcm.6364. https://doi.org/https://doi.org/10.1002/rcm.6364, https://arxiv.org/abs/https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.6364 .

[35] Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **45** (7), 703–714 (2010).

URL https://doi.org/10.1002/jms.1777. https://doi.org/https://doi.org/10.1002/jms.1777 .

[36] NIST. Nist standard reference database 1a (2020). URL https://www.nist.gov/srd/.

[37] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with evo. *bioRxiv* (2024). URL https://www.biorxiv.org/content/early/2024/02/27/2024.02.27.582234. https://doi.org/10.1101/2024.02.27.582234, https://arxiv.org/abs/https://www.biorxiv.org/content/early/2024/02/27/2024.02.27.582234.full.pdf .

[38] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379** (6637), 1123–1130 (2023). URL https://www.science.org/doi/abs/10.1126/science.ade2574. https://doi.org/10.1126/science.ade2574, https://arxiv.org/abs/https://www.science.org/doi/pdf/10.1126/science.ade2574 .

[39] Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* **41** (8), 1099–1106 (2023). URL https://doi.org/10.1038/s41587-022-01618-2. https://doi.org/10.1038/s41587-022-01618-2 .

[40] Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* **381** (6664), eadg7492 (2023). URL https://www.science.org/doi/abs/10.1126/science.adg7492. https://doi.org/10.1126/science.adg7492, https://arxiv.org/abs/https://www.science.org/doi/pdf/10.1126/science.adg7492 .

[41] Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18** (10), 1196–1203 (2021). URL https://doi.org/10.1038/s41592-021-01252-x. https://doi.org/10.1038/s41592-021-01252-x .

[42] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Burstein, J., Doran, C. & Solorio, T. (eds) *BERT: pre-training of deep bidirectional transformers for language understanding.* (eds Burstein, J., Doran, C. & Solorio, T.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, 2019). URL https://doi.org/10.18653/v1/n19-1423.

[43] Brown, T. B. *et al.* Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds) *Language models are few-shot learners.* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) *Advances in Neural*

Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020). URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[44] He, K. *et al.* *Masked autoencoders are scalable vision learners*, 15979–15988 (IEEE, 2022). URL https://doi.org/10.1109/CVPR52688.2022.01553.

[45] Bommasani, R. *et al.* On the opportunities and risks of foundation models. *CoRR* **abs/2108.07258** (2021). URL https://arxiv.org/abs/2108.07258. https://arxiv.org/abs/2108.07258 .

[46] Wang, M. *et al.* Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology* **34** (8), 828–837 (2016). URL https://doi.org/10.1038/nbt.3597. https://doi.org/10.1038/nbt.3597 .

[47] Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 140:1–140:67 (2020). URL http://jmlr.org/papers/v21/20-074.html .

[48] Gemini Team, G. Gemini: A family of highly capable multimodal models (2023). URL https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

[49] Gunasekar, S. *et al.* Textbooks are all you need. *CoRR* **abs/2306.11644** (2023). URL https://doi.org/10.48550/arXiv.2306.11644. https://doi.org/10.48550/ARXIV.2306.11644, https://arxiv.org/abs/2306.11644 .

[50] Singh, A. Mass spectrometry search tool (masst). *Nature Methods* **17** (2), 128–128 (2020). URL https://doi.org/10.1038/s41592-020-0743-x. https://doi.org/10.1038/s41592-020-0743-x .

[51] Quiros-Guerrero, L.-M. *et al.* Inventa: A computational tool to discover structural novelty in natural extracts libraries. *Front Mol Biosci* **9**, 1028334 (2022) .

[52] Velickovic, P. Message passing all the way up. *CoRR* **abs/2202.11097** (2022). URL https://arxiv.org/abs/2202.11097. https://arxiv.org/abs/2202.11097 .

[53] Vaswani, A. *et al.* Guyon, I. *et al.* (eds) *Attention is all you need.* (eds Guyon, I. *et al.*) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008 (2017). URL https://proceedings.neurips.cc/paper/2017/hash/

3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[54] Tancik, M. *et al.* Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds) *Fourier features let networks learn high frequency functions in low dimensional domains.* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020). URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

[55] Kim, S., Rodgers, R. P. & Marshall, A. G. Truly "exact" mass: Elemental composition can be determined uniquely from molecular mass measurement at ∼0.1 mda accuracy for molecules up to ∼500 da. *International Journal of Mass Spectrometry* **251** (2), 260–265 (2006). URL https://www.sciencedirect.com/science/article/pii/S1387380606000856. https://doi.org/https://doi.org/10.1016/j.ijms.2006.02.001, uLTRA-ACCURATE MASS SPECTROMETRY AND RELATED TOPICS Dedicated to H.-J. Kluge on the occasion of his 65th birthday anniversary .

[56] Ying, C. *et al.* Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. & Vaughan, J. W. (eds) *Do transformers really perform badly for graph representation?* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. & Vaughan, J. W.) *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 28877–28888 (2021). URL https://proceedings.neurips.cc/paper/2021/hash/f1c1592588411002af340cbaedd6fc33-Abstract.html.

[57] Alain, G. & Bengio, Y. *Understanding intermediate layers using linear classifier probes* (OpenReview.net, 2017). URL https://openreview.net/forum?id=HJ4-rAVtl.

[58] Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42** (6), 1273–1280 (2002). URL https://doi.org/10.1021/ci010132r. https://doi.org/10.1021/ci010132r .

[59] Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry* **39** (15), 2887–2893 (1996). URL https://doi.org/10.1021/jm9602928. https://doi.org/10.1021/jm9602928 .

[60] McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR* **abs/1802.03426** (2018). URL http://arxiv.org/abs/1802.03426. https://arxiv.org/abs/1802.03426 .

[61] West, K. A., Schmid, R., Gauglitz, J. M., Wang, M. & Dorrestein, P. C. foodmasst a mass spectrometry search tool for foods and beverages. *npj Science of Food* **6** (1), 22 (2022). URL https://doi.org/10.1038/s41538-022-00137-3. https://doi.org/10.1038/s41538-022-00137-3 .

[62] OpenAI. GPT-4 technical report. *CoRR* **abs/2303.08774** (2023). URL https://doi.org/10.48550/arXiv.2303.08774. https://doi.org/10.48550/ARXIV.2303.08774, https://arxiv.org/abs/2303.08774 .

[63] Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **23** (1), 3–25 (1997). URL https://www.sciencedirect.com/science/article/pii/S0169409X96004231. https://doi.org/https://doi.org/10.1016/S0169-409X(96)00423-1, in Vitro Models for Selection of Development Candidates .

[64] Liu, Y., D'Agostino, L. A., Qu, G., Jiang, G. & Martin, J. W. High-resolution mass spectrometry (hrms) methods for nontarget discovery and characterization of poly- and per-fluoroalkyl substances (pfass) in environmental and human samples. *TrAC Trends in Analytical Chemistry* **121**, 115420 (2019). URL https://www.sciencedirect.com/science/article/pii/S0165993618306253. https://doi.org/https://doi.org/10.1016/j.trac.2019.02.021 .

[65] Dong, W., Charikar, M. & Li, K. Srinivasan, S. *et al.* (eds) *Efficient k-nearest neighbor graph construction for generic similarity measures.* (eds Srinivasan, S. *et al.*) *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, 577–586 (ACM, 2011). URL https://doi.org/10.1145/1963405.1963487.

[66] Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **12** (1), 12 (2020). URL https://doi.org/10.1186/s13321-020-0416-x. https://doi.org/10.1186/s13321-020-0416-x .

[67] Mongia, M. *et al.* Fast mass spectrometry search and clustering of untargeted metabolomics data. *Nature Biotechnology* (2024). URL https://doi.org/10.1038/s41587-023-01985-4. https://doi.org/10.1038/s41587-023-01985-4 .

[68] Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W. & Böcker, S. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv* (2023). URL https://www.biorxiv.org/content/early/2023/03/27/2023.03.27.534311. https://doi.org/10.1101/2023.03.27.534311, https://arxiv.org/abs/https://www.

biorxiv.org/content/early/2023/03/27/2023.03.27.534311.full.pdf .

[69] Bittremieux, W. *et al.* Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nature Communications* **14** (1), 8488 (2023). URL https://doi.org/10.1038/s41467-023-44035-y. https://doi.org/10.1038/s41467-023-44035-y .

[70] Griffiths, C. E. M., Armstrong, A. W., Gudjonsson, J. E. & Barker, J. N. W. N. Psoriasis. *The Lancet* **397** (10281), 1301–1315 (2021). URL https://www.sciencedirect.com/science/article/pii/S0140673620325496. https://doi.org/https://doi.org/10.1016/S0140-6736(20)32549-6 .

[71] Hu, W. *et al.* Co-detection of azoxystrobin and thiabendazole fungicides in mold and mildew resistant wallboards and in children. *Heliyon* **10** (6), e27980 (2024). URL https://www.sciencedirect.com/science/article/pii/S2405844024040118. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e27980 .

[72] Haug, K. *et al.* Metabolights: a resource evolving in response to the needs of its scientific community. *Nucleic acids research* **48** (D1), D440–D444 (2020). https://doi.org/10.1093/nar/gkz1019 .

[73] Bushuiev, R. & Pluskal, T. Self-supervised machine learning for the interpretation of molecular mass spectrometry data. URL https://dspace.cvut.cz/handle/10467/108811.

[74] Charikar, M. S. *Similarity estimation techniques from rounding algorithms*, STOC '02, 380–388 (Association for Computing Machinery, New York, NY, USA, 2002). URL https://doi.org/10.1145/509907.509965.

[75] Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation* **5** (2), 107–113 (1965). URL https://doi.org/10.1021/c160017a018. https://doi.org/10.1021/c160017a018 .

[76] Yilmaz, M., Fondrie, W., Bittremieux, W., Oh, S. & Noble, W. S. Chaudhuri, K. *et al.* (eds) *De novo mass spectrometry peptide sequencing with a transformer model.* (eds Chaudhuri, K. *et al.*) *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Vol. 162 of *Proceedings of Machine Learning Research*, 25514–25522 (PMLR, 2022). URL https://proceedings.mlr.press/v162/yilmaz22a.html.

[77] Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges (2021). 2104.13478.

[78] Xiong, R. *et al.* *On layer normalization in the transformer architecture*, Vol. 119 of *Proceedings of Machine Learning Research*, 10524–10533 (PMLR, 2020). URL http://proceedings.mlr.press/v119/xiong20b.html.

[79] Nguyen, T. Q. & Salazar, J. Niehues, J. *et al.* (eds) *Transformers without tears: Improving the normalization of self-attention.* (eds Niehues, J. *et al.*) *Proceedings of the 16th International Conference on Spoken Language Translation, IWSLT 2019, Hong Kong, November 2-3, 2019* (Association for Computational Linguistics, 2019). URL https://aclanthology.org/2019.iwslt-1.17.

[80] Zaheer, M. *et al.* Guyon, I. *et al.* (eds) *Deep sets.* (eds Guyon, I. *et al.*) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 3391–3401 (2017). URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

[81] Zhang, R., Isola, P. & Efros, A. A. *Colorful image colorization*, 649–666 (Springer, 2016). URL https://doi.org/10.48550/arXiv.1603.08511.

[82] Chechik, G., Sharma, V., Shalit, U. & Bengio, S. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.* **11**, 1109–1135 (2010). URL https://dl.acm.org/doi/10.5555/1756006.1756042. https://doi.org/10.5555/1756006.1756042 .

[83] Chen, T. & Guestrin, C. *XGBoost: A scalable tree boosting system*, KDD '16, 785–794 (ACM, New York, NY, USA, 2016). URL http://doi.acm.org/10.1145/2939672.2939785.

[84] Lin, T., Goyal, P., Girshick, R. B., He, K. & Dollár, P. *Focal loss for dense object detection*, 2999–3007 (IEEE Computer Society, 2017). URL https://doi.org/10.1109/ICCV.2017.324.

[85] Kingma, D. P. & Ba, J. Bengio, Y. & LeCun, Y. (eds) *Adam: A method for stochastic optimization.* (eds Bengio, Y. & LeCun, Y.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015). URL http://arxiv.org/abs/1412.6980.

[86] Huber, F. *et al.* matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software* **5** (52), 2411 (2020). URL https://doi.org/10.21105/joss.02411. https://doi.org/10.21105/joss.02411 .

[87] Röst, H. L., Schmitt, U., Aebersold, R. & Malmström, L. pyopenms: A python-based interface to the openms mass-spectrometry algorithm library. *PROTEOMICS* **14** (1), 74–77 (2014). URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201300246. https://doi.org/https://doi.org/10.1002/pmic.201300246, https://arxiv.org/abs/https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201300246 .

[88] Paszke, A. *et al.* Wallach, H. M. *et al.* (eds) *Pytorch: An imperative style, high-performance deep learning library.* (eds Wallach, H. M. *et al.*) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 8024–8035 (2019). URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[89] Falcon, W. Pytorch lightning (2019). URL https://cir.nii.ac.jp/crid/1370013168774120069.

[90] Xu, F. *et al.* Self-supervised EEG representation learning with contrastive predictive coding for post-stroke patients. *Int. J. Neural Syst.* **33** (12), 2350066:1–2350066:16 (2023). URL https://doi.org/10.1142/S0129065723500661. https://doi.org/10.1142/S0129065723500661 .

[91] Song, J., Kim, S. & Yoon, S. Moens, M., Huang, X., Specia, L. & Yih, S. W. (eds) *Alignart: Non-autoregressive neural machine translation by jointly learning to estimate alignment and translate.* (eds Moens, M., Huang, X., Specia, L. & Yih, S. W.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 1–14 (Association for Computational Linguistics, 2021). URL https://doi.org/10.18653/v1/2021.emnlp-main.1.

Online Methods

Construction of GeMS dataset

To enable self-supervised learning, we mine new large datasets of metabolite MS/MS spectra from the GNPS part of MassIVE repository, which we name GeMS (GNPS Experimental Mass Spectra). MassIVE is a community-driven resource with billions of mass spectra from various biochemical and environmental studies. However, it primarily focuses on proteomics and often contains low-quality data as a result of its uncurated nature. Therefore, we have developed a series of algorithms to identify, filter, and cluster the metabolomics spectra of MassIVE into high-quality, non-redundant datasets. In this section, we describe our procedure; a more detailed analysis and statistics are available in our technical report [73].

Selecting LC–MS/MS experiments from MassIVE

We start the mining of MassIVE by selecting all `.mzML` and `.mzXML` data files from all 4,467 MassIVE datasets (as of November 2022) that are explicitly marked as metabolomics studies with the "GNPS" prefix in their names. This selection yields 338,649 distinct files, among which 249,422 contain MS/MS data with a total of 814 million MS/MS spectra. By filtering out empty or corrupted spectra with invalid m/z or intensity values (e.g., negative intensity or multiple identical m/z values), we obtain a complete, unprocessed version of GeMS, comprising 714 million MS/MS spectra.

Estimating quality of MS data

To obtain higher-quality subsets, we apply file-level and spectrum-level quality criteria to the collected spectra. File-level criteria assess the ordering of spectra based on retention times and tandem MS levels. We discard files with unordered retention times, invalid sequences of MS levels (e.g., MS^3 following MS^1 without MS^2), missing MS^1 data, or fewer than three spectra. Importantly, we estimate MS instrument accuracy by evaluating the deviation of similar m/z values within extracted ion chromatograms (XICs). More precisely, the algorithm constructs a set of XICs for MS^1 base peak masses and then estimates the accuracy of the instrument as the median of standard deviations within individual XICs (Algorithm 1).

The spectrum-level quality criteria operate in several steps. Initially, spectra with a low number of peaks or low intensity amplitudes (i.e., the maximum intensity divided by the minimum intensity) are filtered out. Subsequently, non-single charge precursor ions and spectra with excessively high m/z values ($> 1,000$ Da), are excluded. These steps are crucial for retaining only small metabolite molecules. We keep only spectra acquired in positive ionization mode and filter out those estimated to be non-centroided (Algorithm 2).

By varying filtering thresholds, we create three GeMS variants: GeMS A (42 million spectra), GeMS B (100 million spectra), and GeMS C (201 million spectra). GeMS A has a low threshold for estimated instrument accuracy

Algorithm 1 Estimate the absolute accuracy of a mass spectrometry instrument

Require: Sequence of MS^1 spectra from LC–MS experiment.
Ensure: Estimated absolute accuracy of mass spectrometry instrument.

$M_1 \leftarrow$ M/z values of all base peaks ▷ M/z values for 1st round of XICs
$M_2 \leftarrow \{\}$ ▷ M/z values for 2nd round of XICs
for $m \in M_1$ **do**
  $X \leftarrow \mathrm{XIC}(m, 0.5)$ ▷ Set of peaks forming XIC for m/z m and 0.5 Da absolute tolerance
  if $|X| \geq 5$ **then**
    $M_2 \leftarrow M_2 \cup \textsc{MedianMz}(X)$
  end if
end for
$A \leftarrow \{\}$ ▷ Accuracy estimates within individual XICs
for $m \in M_2$ **do**
  $X \leftarrow \mathrm{XIC}(m, 0.01)$ ▷ XIC with lower 0.01 Da tolerance
  if $|X| \geq 5$ **then**
    $A \leftarrow A \cup \textsc{StdDevMz}(X)$
  end if
end for
return $\textsc{Median}(A)$

(approximately four decimal places in m/z ratios). GeMS B is primarily filtered by unknown charge values and is less stringent than GeMS A. GeMS C further relaxes criteria applied to GeMS B and is mainly filtered based on criteria related to spectral peak values. Fig. 2b provides the details of the applied filters for each subset.

Clustering mass spectra with locality-sensitive hashing

The filtering pipeline ensures the quality of individual spectra, but it does not address biases in the entire GeMS datasets related to the natural abundance of metabolites. To tackle this, we employ the random projections algorithm [74] for efficient clustering and deduplication of mass spectra. This algorithm, falling under the family of locality-sensitive hashes, enables linear-time clustering of MS/MS spectra.

In the first step, we vectorize mass spectra via binning. Specifically, each spectrum is represented as a vector $\mathbf{s} \in \mathbb{R}^n$ with n equal-width bins covering the range of m/z values of interest. The value of $\mathbf{s}_i$ then corresponds to the summed intensity of the values contained within the ith bin.

In the subsequent step, for a binned spectrum $\mathbf{s} \in \mathbb{R}^n$, we calculate the corresponding hash $h(\mathbf{s})$ using a mapping $h : \mathbb{R}^n \to \{0, 1\}^m$ defined as

$$h(\mathbf{s}) = [\mathbf{W}\mathbf{s} \geq 0], \text{ where } \mathbf{W} \in \mathbb{R}^{m,n} \text{ and } \mathbf{W}_{ij} \sim \mathcal{N}(0, 1),$$

Algorithm 2 Estimate the type of a spectrum

Require: Spectrum m/z values $\mathbf{m} \in \mathbb{R}^n$ and intensities $\mathbf{i} \in \mathbb{R}^n$.
Ensure: Estimated spectrum type.

if $n < 5$ **then**
  return CENTROID
end if
$b \leftarrow \operatorname{argmax} \mathbf{i}$ ▷ Index of base peak
$S \leftarrow \{s \in \{1, \ldots, n\} \mid (\forall s' \in \{0, \ldots, s-b\})(i_{b+s'} > \frac{i_b}{2})\}$
if $\max S - \min S < 3$ or $\mathbf{m}_{\max S} - \mathbf{m}_{\min S} > \frac{\max \mathbf{m} - \min \mathbf{m}}{1000}$ **then**
  return CENTROID
else
  if $(\exists i \in \mathbf{i})(i = 0)$ **then**
    return PROFILE
  else
    return THRESHOLDED
  end if
end if

where $[\cdot]$ indicates an element-wise Iverson bracket, meaning that $[x_i] = 1$ if x_i is true and 0 otherwise. Essentially, each element of the $\mathbf{Ws}$ product is a dot product of $\mathbf{s}$ and a random n-dimensional hyperplane. Each of the m hyperplanes splits the n-dimensional space into two complementary subspaces, thereby determining the subspace to which $\mathbf{s}$ belongs, based on the sign of each dot product. These signs represent the bits of the resulting m-dimensional hash. Given that every hyperplane intersects the origin, the likelihood of two binned spectra $\mathbf{s}_i$ and $\mathbf{s}_j$ sharing the same hash is a function of their cosine similarity [74]:

$$\mathbb{P}(h(\mathbf{s}_i) = h(\mathbf{s}_j)) = 1 - \arccos\left(\underbrace{\frac{\mathbf{s}_i^\top \mathbf{s}_j}{\|\mathbf{s}_i\|\|\mathbf{s}_j\|}}_{\text{Cosine similarity}}\right)\frac{1}{\pi}, \tag{1}$$

where $\mathbb{P}$ denotes the joint probability over random hyperplanes. In essence, with a sufficient number of hyperplanes, random projections effectively approximate cosine similarity, which is the primary method for comparing mass spectra.

To cluster the spectra of GeMS, we use $m = 1,000$ random hyperplanes and the window of size 1 binning the range of m/z values from 0 to 1,000 Da (i.e., $n = 1,000$). By varying the number of retained spectra per cluster, we establish two additional subsets for each of the A, B, and C variants of GeMS with at most 10 and 1,000 allowed cluster representatives, denoted with additional suffixes such as GeMS-A1 or GeMS-B1000. Fig. 2c demonstrates the sizes of the resulting clustered datasets.

GeMS data format

We store GeMS datasets in a compressed tensor format using our new `.hdf5`-based format, primarily designed for deep learning. Extended Data Table 4 outlines the format specifications, detailing all data and metadata entities retained from the input `.mzML` or `.mzXML` files.

Murcko histograms algorithm for splitting molecular datasets

A universal and reliable protocol for supervised learning on spectral libraries is crucial for fine-tuning our pre-trained DreaMS model. The commonly used technique is to split a spectral library into training and validation folds, ensuring no molecules share identical structures (technically, the first 14 characters of InChI keys) between the folds. However, we identify three issues with this protocol which may limit the generalization capabilities and, therefore, the practical utility of the final fine-tuned model.

First, spectral libraries often contain closely similar structures [73, Section 4.1], such as those resulting from click chemistry. Consequently, molecules with minor structural differences are often assigned to different train-validation folds, introducing a data leakage for tasks such as fingerprint prediction, where small structural details may not significantly impact performance metrics. Second, structure-disjoint splits are agnostic to the fragmentation nature of tandem mass spectrometry. For instance, two molecules differing only in the length of the carbon chain connecting two subfragments have distinct structures, yet such chains can be easily fragmented by CID, resulting in nearly identical spectra. Third, the structure-disjoint approach often assigns entire molecules and their abundant fragments (such as the fragments of sugars) to different folds, increasing the chance of overfitting to abundant substructures. To address these issues, we have designed a new algorithm, Murcko histograms, based on the Murcko scaffolds [59], for splitting molecular structures into training-validation folds.

To address the first issue, we build our method upon coarse-grained Murcko scaffolds. To tackle the second issue of insensitivity to fragmentation, our method operates on molecular fragments as the primary design principle. To address the third issue, we define a heavily relaxed notion of molecular similarity, ensuring that the distinction between folds is well-defined.

In particular, our algorithm computes a histogram defined in terms of the counts of scaffold substructures (Algorithm 3). Given the Murcko scaffold of a molecule [59], the algorithm operates on two separate groups of its atoms. The first group consists of sets of atoms, with each set determining a ring (line 2 in the algorithm), whereas the second group includes all atoms connecting these rings (i.e., linkers; line 3). For each ring, the algorithm calculates a pair of natural numbers: the number of neighboring rings and the number of adjacent linkers (denoted as r, l in lines 5–9). These pairs define the domain of the resulting histogram, where the values represent the counts of such pairs within

Algorithm 3 Definition of a Murcko histogram

Require: Molecular graph $G = (V, E), V = \{1, \ldots, n\}, E \subseteq \{\{u, v\} \mid u, v \in V \wedge u \neq v\}$.

Ensure: Murcko histogram h.

1: $G \leftarrow$ MurckoScaffold(G)
2: $V_R \leftarrow \{V_r \subset V \wedge |V_r| > 3 \mid V_r$ contains all atoms of a (fused) ring$\}$
3: $V_L \leftarrow \{v \in V \mid \deg(v) > 1 \wedge v$ is not in any ring$\}$
4: $h \leftarrow$ a map $\mathbb{N}^2 \to \mathbb{N}$ initialized as $(\forall i, j \in \mathbb{N}^2)(h(i, j) = 0)$
5: **for** $V_r \in V_R$ **do**
6: $\quad r \leftarrow \sum\{|V_r \cap V_r^{'}/2| \mid V_r^{'} \in V_R \setminus V_r\}$
7: $\quad l \leftarrow |V_r \cap V_l|$
8: $\quad h(r, l) \leftarrow h(r, l) + 1$
9: **end for**
10: **return** h

a molecule (lines 4, 10). Extended Data Fig. 1a shows examples of Murcko histograms and the corresponding molecular structures.

The Murcko histogram-disjoint train-validation splitting resolves the first two aforementioned issues by being insensitive to minor atomic details and by taking into account the fragments of molecular scaffolds instead. We further address the third issue by defining a way to compare the histograms which is more relaxed than a simple identity (Algorithm 4). Specifically, we define a distance on Murcko histograms as the difference in the histogram values solely in rings, not considering the number of neighboring linkers. Using this definition, we relocate the samples from validation to train folds if their distance is less than 5, while not performing the relocation if the minimum number of rings in one of the molecules is less than 4. Notice that these parameters provide interpretability for the boundary between train and validation folds, and by varying them, we can balance between the number of validation examples and the degree of similarity between train and validation folds in terms of scaffold substructures.

Unlike structure-disjoint splitting, our method eliminates virtually all near-duplicate training-validation leaks, resulting in a two-fold reduction in average Morgan Tanimoto similarity [75] between the molecules corresponding to training and validation spectra (Extended Data Fig. 1b).

With this approach, we define approximately 90%/10% training-validation splits for MoNA as well as the union of MoNA and NIST20, which we use for fine-tuning. Throughout the text, we refer to these splits as the Murcko histogram-disjoint splits. As mentioned previously, the name originates from the use of Murcko scaffolds [59] as the basis for the algorithm. We anticipate that our training-evaluation protocol based on Murcko histograms will stimulate further research into the development of a new generation of models with enhanced generalization towards the undiscovered dark metabolome [5].

Algorithm 4 Definition of a Murcko subhistogram relation

Require: Two Murcko histograms h_1 and h_2, a minimum number of rings k to compute the non-identity relation, and a minimum difference in ring-only Murcko histogram m to consider the histograms different. The default values are $k = 4$ and $m = 5$.

Ensure: TRUE if one of h_1, h_2 is a subhistogram of the other in Murcko rings, FALSE otherwise.

if $\min\{\sum_{i,j\in\mathbb{N}} h_1(i,j), \sum_{i,j\in\mathbb{N}} h_2(i,j)\} < k$ **then**
    return $h_1 = h_2$
end if
$d \leftarrow \sum_{i\in\mathbb{N}} |(\sum_{j\in\mathbb{N}} h_1(i,j) - \sum_{j\in\mathbb{N}} h_2(i,j))|$
if $d < m$ **then**
    return TRUE
else
    return FALSE
end if

DreaMS neural network architecture

The DreaMS neural network architecture (Fig. 3b) can be decomposed into three main consecutive modules. Given a mass spectrum, the model first encodes each spectral peak into a high-dimensional continuous representation with PEAKENCODER. Then, it processes the entire set of encoded peaks with SPECTRUMENCODER – a series of transformer encoder blocks [53]. Each block learns relationships between peaks and consecutively enriches their representations. The final task-specific PEAKDECODER adjusts the final transformer representations according to a task-specific training objective. Each of the modules is described in detail below.

PeakEncoder

We represent each raw mass spectrum as a matrix $\mathbf{S} \in \mathbb{R}^{2,n+1}$, constructed as

$$\mathbf{S} = \begin{bmatrix} m_0 & m_1 & m_2 & \dots & m_n \\ 1.1 & i_1 & i_2 & \dots & i_n \end{bmatrix}, \tag{2}$$

where each column, indexed by $j \in \{1, \dots, n\}$, corresponds to one of the n spectral peaks and is represented as the continuous vector $[m_j, i_j]^\top \in \mathbb{R} \times [0, 1]$, denoting the pair of m/z and relative intensity values (m/z denoted by m and intensity denoted by i). Additionally, we prepend a precursor m/z m_0 and assign it an artificial intensity of 1.1. We term this additional peak the precursor token and utilize it as a master node [52] for aggregating spectrum-level information. If a spectrum has more than n peaks, we select the n most intense ones; if it has fewer than n peaks, we pad the matrix $\mathbf{S}$ with zeros.

Rather than treating each m/z ratio as a single continuous value, we process it using a mass-tolerant modification of Fourier features $\Phi : \mathbb{R} \to [-1, 1]^{2B}$ [54],

dependent on B predefined frequencies $\mathbf{b} \in \mathbb{R}^B$. Specifically, the features are constructed with sine and cosine functions

$$\Phi(m)_i = \sin(2\pi b_i m), \qquad \Phi(m)_{i+1} = \cos(2\pi b_i m), \tag{3}$$

where each frequency b_i is uniquely associated with either a low frequency capturing the integer part of a mass $m \in \mathbb{R}$ or a high frequency capturing its decimal part, forming together a vector of frequencies

$$\mathbf{b} = \Big[\underbrace{\frac{1}{m_{max}}, \frac{1}{m_{max}-1}, \ldots,}_{\text{Low frequencies}} \frac{1}{1}, \underbrace{\frac{1}{km_{min}}, \frac{1}{(k-1)m_{min}}, \ldots, \frac{1}{m_{min}}}_{\text{High frequencies}} \Big]^\top \in \mathbb{R}^B. \tag{4}$$

Here, constants $m_{\text{min}} \in (0, 1)$ and $m_{\text{max}} \in (1, \infty)$ represent the minimum decimal mass of intereset (i.e., the absolute instrument accuracy) and the maximum integer mass of interest, and $k \in \mathbb{N}$ is such that km_{min} is the closest value to 1. For instance, when training DreaMS on GeMS-A spectra, we set $m_{\text{min}} = 10^{-4}$ and $m_{\text{max}} = 1000$ according to the construction of GeMS datasets. This schema yields 1000 low frequencies and 5000 high frequencies (i.e., the overall dimensionality of the vector $\mathbf{b}$ is 6000.).

Further, we process the Fourier features given by Equation 3 with a feed-forward neural network $\text{FFN}_F : \mathbb{R}^{2B} \to \mathbb{R}^{d_m}$. We hypothesize that the sensitivity of Fourier features to both large and small differences in masses allows FFN_F to learn the space of plausible molecular masses given by elemental compositions. Our instantiation of frequencies outperforms both random initialization [54] and the log-spaced sinusoidal variant proposed for proteomics [14, 76] (Extended Data Fig. 4; [73]). Notably, since peaks do not form a sequence of tokens but rather a set, we do not encode their positions, in contrast with classic positional encoding [53].

The concatenation of the output of FFN_F with the output of another shallow feed-forward network $\text{FFN}_P : \mathbb{R}^2 \to \mathbb{R}^{d_p}$ applied to raw m/z and intensity values forms the complete $\textsc{PeakEncoder} : \mathbb{R}^2 \to \mathbb{R}^{d_m+d_p}$:

$$\textsc{PeakEncoder}(m, i) = \text{FFN}_F(\Phi(m)) \parallel \text{FFN}_P(m, i), \tag{5}$$

where $\parallel$ denotes concatenation. Column-wise application of $\textsc{PeakEncoder}$ to the matrix $\mathbf{S}$ yields a high-dimensional representation of the corresponding spectrum $\mathbf{S}_0 \in \mathbb{R}^{d,n}$, where $d = d_m + d_p$ is the dimensionality of the representation and n is the number of peaks.

SpectrumEncoder

Given the output of $\textsc{PeakEncoder}$, $\textsc{SpectrumEncoder} : \mathbb{R}^{d,n} \to \mathbb{R}^{d,n}$ updates the representations of peaks by exchanging information between individual peaks via the self-attention mechanism. This is achieved through a

sequence of l transformer encoder layers (i.e., BERT [42]), alternating multi-head self-attention blocks with peak-wise feed-forward networks. Starting from $\mathbf{S}_0$, each i-th block gradually updates the representation of the spectrum from $\mathbf{S}_{i-1}$ to $\mathbf{S}_i$. Throughout the text, we denote the columns of $\mathbf{S}_l$ (i.e., representations of individual peaks) as $\mathbf{s}_0, \ldots, \mathbf{s}_n$. We refer to the first columns of such matrices, representing precursor tokens, as DreaMS (Deep Representations Empowering the Annotation of Mass Spectra).

An important property of the transformer encoder is its equivariance to permutations of tokens [77]. Combined with the position-invariant encoding of peaks through PeakEncoder, this implies that the same two peaks in different spectra will have identical attention scores in the first attention layer, regardless of the total number of peaks or noise signals between these two peaks. To further strengthen the inductive bias of the transformer towards the relations between peaks, we explicitly enrich the attention mechanism with all pairwise m/z differences including neutral losses. In each transformer layer, the attention score $\mathbf{A}_{ij}$ between the i-th and j-th peaks is computed as:

$$\mathbf{A}_{ij} = \frac{\mathbf{q}_i^\top \mathbf{k}_j + \sum_k^{2t} \Phi(m_i)_k - \Phi(m_j)_k}{\sqrt{d}}, \tag{6}$$

where $\mathbf{q}_i^\top \mathbf{k}_j$ is a standard dot-product attention and $\sum_k^{2t} \Phi(m_i)_k - \Phi(m_j)_k$ is an additional Graphormer-like term [56]. Element-wise differences in Fourier features enable the transformer to directly attend to precise m/z differences, enhancing its capacity to learn fragmentation patterns and robustness to shifts in absolute m/z values. This is particularly important, for instance, in scenarios where m/z values are shifted due to the masses of ionization adducts.

In contrast to BERT, we use a pre-norm variant of transformer [78], remove biases in linear layers, and use ReLU activations. We utilize the implementation of transformer provided by Nguyen et al. [79].

PeakDecoder

Depending on the training objective, we use linear layers of different shapes (referred to as heads) to refine and project the final hidden representations of peaks given by the SpectrumEncoder.

For both m/z masking and retention order pre-training objectives, we employ simple linear projections followed by suitable activation functions, mapping the representations of peaks into the corresponding domains of predictions:

$$\hat{\mathbf{y}}_{\text{mass}} = \text{softmax}(\mathbf{W}_{\text{mass}} \mathbf{s}_k), \qquad \hat{y}_{\text{order}} = \sigma(\mathbf{W}_{\text{order}} (\mathbf{s}_0^{(i)} \| \mathbf{s}_0^{(j)})), \tag{7}$$

where $\mathbf{s}_k \in \mathbb{R}^d$ denotes the hidden representation of a masked peak $k \in M$ from a set of masked indices $M \subset \{1, \ldots, n\}$. It is projected by $\mathbf{W}_{\text{mass}} \in \mathbb{R}^{c,d}$ and the softmax function to obtain the predicted probability vector $\hat{\mathbf{y}}_{\text{mass}} \in \mathbb{R}^c$ with c classes corresponding to the discretized mass bins to be reconstructed.

Next, $\hat{y}_{\text{order}}$ denotes the predicted probability that a spectrum i precedes the spectrum j in chromatography. The probability is predicted by concatenating two precursor embeddings $\mathbf{s}_0^{(i)}, \mathbf{s}_0^{(j)}$corresponding to the two spectra, and applying the linear projection $\mathbf{W}_{\text{order}} \in \mathbb{R}^{1,2d}$ followed by the sigmoid function σ.

For supervised fine-tuning tasks, we emloy two variants of linear heads. The first variant is given by single linear layers operating solely on the precursor token representations $\mathbf{s}_0 \in \mathbb{R}^d$:

$$\hat{\mathbf{y}}_{\text{props}} = \mathbf{W}_{\text{props}}\mathbf{s}_0, \qquad \hat{y}_{\text{F}} = \sigma(\mathbf{W}_{\text{F}}\mathbf{s}_0), \qquad \mathbf{z} = \mathbf{W}_{\text{emb}}\mathbf{s}_0, \tag{8}$$

where $\mathbf{W}_{\text{props}} \in \mathbb{R}^{11,d}$, $\mathbf{W}_{\text{F}} \in \mathbb{R}^{1,d}$ followed by sigmoid σ, and $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{d,d}$ yield the predictions of eleven molecular properties $\hat{\mathbf{y}}_{\text{props}}$, the probability of fluorine presence $\hat{y}_{\text{F}}$, and the spectral embedding $\mathbf{z}$, respectively.

For the task of predicting molecular fingerprints, we find a head with richer representation capacity to slightly improve the performance:

$$\hat{\mathbf{y}}_{\text{fp}} = \mathbf{W}_{\text{fp1}} \sum_{i=0}^{n} \text{ReLU}(\mathbf{W}_{\text{fp0}}\mathbf{s}_i). \tag{9}$$

Here, the projections $\mathbf{W}_{\text{fp0}} \in \mathbb{R}^{d,d}$ and $\mathbf{W}_{\text{fp1}} \in \mathbb{R}^{4096,d}$ are arranged into the DeepSets-like [80] head to output 4096 fingerprint elements. In this case, the head operates on the hidden representations of all peaks $\mathbf{s}_i$ rather than solely on the precursor peak $\mathbf{s}_0$ as in Equation 8. The details of pre-training and fine-tuning objectives are discussed in the following sections.

Self-supervised pre-training

The objective of self-supervised pre-training for DreaMS is defined by minimizing a weighted sum of two losses:

$$\mathcal{L}_{\text{DreaMS}} = 0.8\mathcal{L}_{\text{mass}} + 0.2\mathcal{L}_{\text{order}}, \tag{10}$$

where $\mathcal{L}_{\text{mass}}$ represents the masked modeling loss, quantifying the error of the model in reconstructing the masses of randomly masked peaks, and $\mathcal{L}_{\text{order}}$ denotes the retention order prediction error. Each training example within a mini-batch consists of sampling two spectra with indices i, j from the same LC–MS/MS experiment. Here, we further detail the computation of both $\mathcal{L}_{\text{mass}}$ and $\mathcal{L}_{\text{order}}$ losses for the example pair i, j.

To compute the $\mathcal{L}_{\text{mass}}$ loss, we randomly sample a predefined ratio of peaks $M^{(i)}, M^{(j)} \subset \{1, \ldots, n\}$ from both spectra i and j, proportionally to their intensities. Then, we replace the masses of the sampled peaks in the spectra with -1.0, while keeping the intensities unchanged, and utilize the original mass values $\mathbf{m}^{(i)} \in \mathbb{R}^{|M^{(i)}|}$ and $\mathbf{m}^{(j)} \in \mathbb{R}^{|M^{(j)}|}$ as the prediction labels. Instead of directly predicting the continuous values $\mathbf{m}^{(i)}, \mathbf{m}^{(j)}$, we categorize them

into c equal-width bins ranging from 0 to the maximum m/z of the training dataset (1000 Da for GeMS-A subsets; Fig. 2b) and train the model to predict the correct bins. This classification approach [81], rather than regression, is adopted to better capture the inherent uncertainty of mass reconstruction, as it accounts for the possibility that several masses may be equally plausible for a masked peak. A regression model may converge at predicting the average value whereas a classification model would learn to assign equal probability to each plausible mass.

Specifically, we convert continuous mass values into degenerate categorical distributions, represented by binary matrices $\mathbf{Y}^{(i)}_{mass} \in \{0,1\}^{|M^{(i)}|,c}$ and $\mathbf{Y}^{(j)}_{mass} \in \{0,1\}^{|M^{(j)}|,c}$, where rows correspond to masked peaks and columns correspond to mass bins. The elements of the matrices are ones in bins containing the corresponding masses and zeros elsewhere. In detail, for a masked peak $l \in M^{(k)}$ in spectrum $k \in \{i,j\}$ and bin $b \in \{0,\dots,c-1\}$, the corresponding matrix element is

$$y^{(k)}_{\text{mass},l,b} = \left[m^{(k)}_l \in [b\frac{1000}{c}, (b+1)\frac{1000}{c})\right], \tag{11}$$

where $[\cdot]$ indicates the Iverson bracket, implying $[x] = 1$ if x is true and 0 otherwise. The terms $\frac{1000}{c}$ represent the m/z range $(0, 1000)$ discretized into c bins.

Then, the model is trained to predict a categorical distribution for each of the masked peaks $\hat{\mathbf{Y}}^{(i)}_{\text{mass}}, \hat{\mathbf{Y}}^{(j)}_{\text{mass}}$ (Equation 7, left) and the reconstruction error is evaluated using the cross-entropy loss in the space of discretized mass values:

$$\mathcal{L}_{\text{mass}}(\hat{\mathbf{Y}}^{(i)}_{\text{mass}}, \mathbf{Y}^{(i)}_{\text{mass}}, \hat{\mathbf{Y}}^{(j)}_{\text{mass}}, \mathbf{Y}^{(j)}_{\text{mass}}) = -\frac{1}{2}\sum_{k\in\{i,j\}}\sum_{l\in M^{(k)}} {\mathbf{y}^{(k)}_{\text{mass},l}}^{\top} \log(\hat{\mathbf{y}}^{(k)}_{\text{mass},l}), \tag{12}$$

where the first sum from the left averages the results across two sampled spectra i and j, and the second sum iterates over all masked peaks $M^{(k)}$ in spectrum k. The dot product ${\mathbf{y}^{(k)}_{\text{mass},l}}^{\top} \log(\hat{\mathbf{y}}^{(k)}_{\text{mass},l})$ calculates the cross-entropy between a ground-truth degenerate distribution $\mathbf{y}^{(k)}_{\text{mass},l}$, which contains a one for the correct mass bin of peak l in spectrum k and zeros elsewhere, and the corresponding predicted distribution over bins $\hat{\mathbf{y}}^{(k)}_{\text{mass},l}$. Minimizing $\mathcal{L}_{mass}$ effectively maximizes the likelihood of predicting the correct mass bins, and the loss is minimal when all the bins are predicted correctly.

The second component of the DreaMS loss, $\mathcal{L}_{\text{order}}$, is given by a binary cross-entropy classification loss. The model is trained to predict the retention order of two spectra i and j within the LC–MS/MS experiment by estimating the probability $\hat{y}_{\text{order}}$ that spectrum i precedes spectrum j in chromatography (Equation 7, right). The actual probability y_{order} is either 0 or 1:

$$\mathcal{L}_{\text{order}} = -(y_{\text{order}} \log(\hat{y}_{\text{order}}) + (1 - y_{\text{order}}) \log(1 - \hat{y}_{\text{order}})). \tag{13}$$

We pre-train DreaMS on the GeMS-A10 dataset and retain the sixty highest peaks when forming training batches. Additionally, with a 20% probability, we augment a spectrum by adding a random scalar from $(0, 50)$ to all its m/z values. Such modification forces the neural network to learn relationships between spectral peaks rather than memorizing precise masses, a property important for making the model more robust to, for example, different ionization adducts.

Linear probing of the emergence of molecular structures

Every 2,500 pre-training iterations, we conduct linear probing – a technique enabling us to evaluate the gradual emergence of molecular structures during self-supervision. Specifically, we freeze a model and train a single linear layer $\mathbf{W}_{\text{probe}} \in \mathbb{R}^{166,d}$ to predict 166 MACCS fingerprint bits from precursor token embeddings, utilizing a random subsample of 6,000 examples from the Murcko histogram split of NIST20 and MoNA. We employ a binary cross-entropy loss function (Equation 13) for learning individual fingerprint bits. We select MACCS fingerprints as the probing objective because they offer an interpretable description of a molecular structure, allowing each predicted bit to be reconstructed back to a molecular substructure.

We report the average validation recall in predicted bits as a function of pre-training time (Fig. 3c) to illustrate the model's progressively improving ability to discover the substructures of ground truth molecules. For each iteration, we display the highest recall within 100 probing epochs. Notably, although the figure depicts only the increase in recall, this improvement is achieved without any decline in precision. In fact, precision slightly increases from 0.81 to 0.84 within the same evaluation setup.

Transfer learning to spectrum annotation tasks

In this section, we discuss how we transfer the knowledge obtained by the DreaMS model during the self-supervised pre-training to make predictions in scenarios of practical interest. Specifically, we describe how we fine-tune the architecture for different downstream mass spectrometry tasks with task-specific heads.

Spectral similarity

The cosine similarity on unsupervised DreaMS embeddings exhibits a strong correlation with Tanimoto similarity (Fig. 4a). However, we observe that it lacks sensitivity to small structural differences among molecules with nearly identical masses (Extended Data Fig. 3b). To address this limitation, we refine the embedding space through contrastive fine-tuning. Specifically, we utilize triplet margin loss function [82] to disentangle the embeddings of spectra which share similar molecular masses:

$$\mathcal{L}_{\text{emb}}(\mathbf{z}, \mathbf{z}^+, \mathbf{z}^-) = \max\{\cos(\mathbf{z}, \mathbf{z}^+) - \cos(\mathbf{z}, \mathbf{z}^-) + \Delta, 0\}, \tag{14}$$

where $\mathbf{z} \in \mathbb{R}^d$ denotes the embedding of a randomly sampled reference spectrum, $\mathbf{z}^+$, $\mathbf{z}^- \in \mathbb{R}^d$ are the embeddings of positive and negative examples, respectively, and $\Delta > 0$ is the contrastive margin. The positive example is defined as a spectrum of the same molecule as the reference spectrum (having the same 14-character prefix in the InChI key) whereas the negative example is given by a spectrum corresponding to a different molecule but with a similar molecular mass (at most 0.05 Da difference). The $\mathcal{L}_{emb}$ loss function optimizes the embedding space so that the reference spectra are closer to the positive examples than to the negative ones. The contrastive margin Δ, intuitively, measures the minimum required gap between the corresponding positive and negative distances. The proximity between two embeddings $\mathbf{a}$ and $\mathbf{b}$ is measured by cosine similarity:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\max\{\|\mathbf{a}\|\|\mathbf{b}\|, \epsilon\}}, \tag{15}$$

where ϵ, set to 10^{-8}, is a constant for numerical stability.

The aim of the fine-tuning is to adjust the embedding space using minimal supervision, yet still retaining the knowledge acquired during self-supervised pre-training and not introducing biases of spectral libraries scarcity. Therefore, we conduct contrastive training on a refined subset of MoNA histogram-disjoint split containing approximately 25,000 spectra corresponding to 5,500 unique InChI connectivity blocks and do not use any spectra from NIST20 for training. To form a subset, we retain only the spectra satisfying A quality conditions (as shown in Fig. 2b), having [M+H]+ adducts and 60 eV collision energy. To simulate the performance evaluation on a new spectral library, we evaluate the cosine similarity in refined embedding space on the high-quality subset of NIST20 satisfying A filtering conditions. We additionally exclude from the validation all NIST20 examples whose InChI key connectivity blocks are present in MoNA. We consider two molecular similarity tasks: estimating the Tanimoto similarity between Morgan fingerprints of underlying molecules, and determining the spectra corresponding to the same molecules within the pool of candidate spectra with similar precursor masses.

Specifically, in the case of the Tanimoto similarity approximation problem, we measure Pearson correlation between DreaMS cosine similarities and Tanimoto similarities on binary Morgan fingerprints (number of bits = 4096, radius = 2) using approximately 82,000 pairs of spectra sampled from NIST20 so that they maximize the entropy of the distribution of ground-truth similarities. We benchmark our method against the official implementation (https://github.com/matchms/ms2deepscore) of the state-of-the-art MS2DeepScore model [13] (as depicted in Fig. 4a).

For the second task of retrieving mass spectra corresponding to the same molecule, we measure the area under the receiver operating characteristic curve (AUROC), which evaluates the classification performance under different similarity thresholds. We sample approximately 750,000 binary classification

examples from NIST20 in a way that makes positive class examples correspond to pairs of spectra having the same underlying molecular structure (as measured by the same 14-character prefix in the InChI key) and negative class examples correspond to pairs of spectra having similar precursor masses (with at most 10 ppm precursor m/z difference). We benchmark our method against spectral entropy, the state-of-the-art method, as well as 43 other baseline approaches [9] (as illustrated in Fig. 4b,c). We use the implementation of all the methods from the official spectral entropy GitHub repository (https://github.com/YuanyueLi/SpectralEntropy).

For the visualization of fine-tuned embeddings (Fig. 4d, Fig. 5, Extended Data Fig. 2), we utilize the UMAP algorithm [60], with cosine similarity set as the metric. Fig. 4d and Fig. 2 display 100,000 random embeddings of NIST20 spectra, with all precursor InChI keys disjoint from the precursors of the MoNA subset used for the spectral similarity fine-tuning. Level set plots in Fig. 2 present ten levels of various molecular properties when binning the UMAP axes into 200 bins. Sample-average embeddings in Fig. 5 are computed for 2,810 food samples (i.e., `.mzML` files; 6 million spectra in total) from the MSV00008490 dataset (https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=ce3254fe529d43f48077d7ad55b7da09), which have textual food descriptions assigned in the metadata table within the dataset repository. We allocate color categories to individual points by querying ChatGPT 4 [62] to summarize all the individual textual descriptions into a minimal number of categories, including the "Miscellaneous" category. We drop 948 samples forming this category, such as the ones containing "supplement" or "extract" in their descriptions.

Molecular fingerprint prediction

The next problem we tackle with DreaMS is the prediction of molecular fingerprints. We adapt our model via supervised fine-tuning and validate it on the MIST CANOPUS benchmark [31] to evaluate the performance against the state-of-the-art model MIST.

In detail, we fine-tune DreaMS to directly predict molecular fingerprints via the cosine similarity loss function $\mathcal{L}_{\text{fp}}$ between true $\mathbf{y}_{\text{fp}}$ and predicted $\hat{\mathbf{y}}_{\text{fp}}$ fingerprints:

$$\mathcal{L}_{\text{fp}}(\hat{\mathbf{y}}_{\text{fp}}, \mathbf{y}_{\text{fp}}) = \cos(\hat{\mathbf{y}}_{\text{fp}}, \mathbf{y}_{\text{fp}}), \tag{16}$$

where cos is the cosine similarity given by Equation (15), discussed previously in the context of comparing embeddings of spectra.

For the fine-tuning and evaluation, we use the CANOPUS benchmark from the official GitHub repository (https://github.com/samgoldman97/mist). Specifically, we use the MIST codebase to generate fingerprints and the candidate pools of molecules for the evaluation. Each pool corresponds to a single spectrum along with positive and negative candidate molecules mined from PubChem. The positive candidates correspond to molecules in

PubChem that have the same 14-character prefix in the InChI key as the true underlying molecule, including the true molecule itself. The negative candidates are given by the molecules sharing the same molecular formula. Then, the retrieval performance is evaluated using the accuracy at top k metrics for $k \in \{1, 5, 10, 20, 50, 100, 200\}$, measuring the number of spectra that have at least one positive molecule in the top k predictions, sorted by the cosine similarity between the predicted and ground-truth fingerprints (Fig. 4e, Extended Data Table 1).

Molecular property prediction

Next, we fine-tune DreaMS to predict molecular properties. For this, we reproduce the evaluation protocols proposed previously by Voronov et al. [25] and Gebhard et al. [26].

Specifically, we fine-tune our model to jointly predict $r = 11$ selected molecular properties from spectra, averaging the squared error for each of the properties:

$$\mathcal{L}_{\text{props}}(\hat{\mathbf{y}}_{\text{props}}, \mathbf{y}_{\text{props}}) = \frac{1}{r}\|\hat{\mathbf{y}}_{\text{props}} - \mathbf{y}_{\text{props}}\|^2, \tag{17}$$

where $\mathbf{y}_{\text{props}} \in \mathbb{R}^r$ denotes the vector containing ground-truth molecular properties, such as quantitative estimation of drug-likeness (QED), synthetic accessibility, and Bertz complexity, (Fig. 4 for the complete list). Because different properties have different scales and are measured in different units, we normalize them before feeding them to the loss function. In particular, we map each property to the $[0, 1]$ interval via min-max scaling based on the statistics from the training data.

For the training, validation and testing, we use the MoNA and NIST20 dataset splits prepared using our Murcko histograms algorithm. First, inspired by Gebhard et al. [26], we evaluate the performance of DreaMS on predicting molecular complexity from mass spectra. In detail, we estimate the capability of DreaMS to predict the Bertz complexity of a molecule from its mass spectum, by measuring its relative prediction error under different minimum true complexity thresholds of interest. The relative prediction error is defined as $|y_{\text{Bertz}} - \hat{y}_{\text{Bertz}}|/y_{\text{Bertz}}$, and measures the performance of predicting complexity $\hat{y}_{\text{Bertz}}$ robustly under varying absolute values of the true complexity y_{Bertz} [26]. We compare our method against XGBoost [26, 83] trained on 1000-dimensional binned spectra with 0.1 Da bin size and the state-of-the-art spectra property predictor MS2Prop [25], reimplemented and retrained to predict Bertz complexity among other properties (Fig. 4f). We also evaluate our method and XGBoost on predicting ten other properties addressed by MS2Prop (Fig. 4g). Our reimplementation of MS2Prop uses the hyperparameters described in the original publication [25] and the same values as DreaMS for the unspecified hyperparameters (such as batch size and learning rate).

Fluorine detection

We evaluate the performance of DreaMS on detecting fluorinated molecules from mass spectra.

Our fluorine detector is fine-tuned using a binary cross entropy loss function $\mathcal{L}_{\text{F}}$ with additional focal loss terms [84] accounting for class imbalance. For each training example, the loss is computed as:

$$\mathcal{L}_{\text{F}}(\hat{y}_{\text{F}}, y_{\text{F}}) = -\alpha_F (1 - p_{\text{F}})^{\gamma} \log p_{\text{F}}, \tag{18}$$

where $\hat{y}_{\text{F}}$ is the predicted fluorine presence probability and y_{F} is the 0 or 1 label, depending on the ground-truth presence of fluorine. Next, p_F is the standard binary cross entropy term, and α_F and γ are focal loss terms:

$$p_F = \begin{cases} \hat{y}_{\text{F}}, & \text{if } y_{\text{F}} = 1 \\ 1 - \hat{y}_{\text{F}}, & \text{otherwise,} \end{cases} \qquad \alpha_{\text{F}} = \begin{cases} \alpha, & \text{if } y_{\text{F}} = 1 \\ 1 - \alpha, & \text{otherwise,} \end{cases} \tag{19}$$

where $\alpha = 0.8$ increases the loss for underrepresented examples, containing fluorine, and decreases the loss otherwise (training data contains approximately 80% of examples with fluorine); $\gamma = 0.5$ adjusts the predicted probabilities of correct classes to prioritize misclassified examples.

We fine-tune DreaMS on the spectra from MoNA and NIST using the Murcko histograms algorithm for training-validation splitting. Subsequently, we test the performance of the model on our in-house dataset, consisting of 17,052 [M+H]+ Orbitrap mass spectra (3,900 spectra of 1,175 unique fluorinated molecules and 13,152 spectra of 4,055 unique non-fluorinated molecules), by measuring precision and recall under different thresholds (Fig. 4h). The dataset is available under the MassIVE accession number MSV000094528 (https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=676a38e2dd574a15905e807d78cf1e57). As a baseline, we use SIRIUS 5.6.3 with the possible adducts set to [M+H]+, the instrument to Orbitrap, and the maximum number of fluorine elements to 5 (maximum number in the dataset) [29]. By experimenting with the numbers lower than 5, we observe a significant drop in recall but no improvement in precision.

We prioritize high precision over recall as we find it the most practically important metric when searching for new fluorinated molecules for further wet-lab characterization, considering the difficulty of wet-lab experiments. Consequently, we estimate the coverage of mass spectra with confident predictions using the model operating in the high-precision regime with the precision of 90%. Specifically, we set two predicted probability thresholds (0.48 and 0.78) for classifying spectra containing and not containing fluorine so as to lend the model 90% precision in both cases. Notably, we find that only 5% of spectra have uncertain predictions (with the predicted probabilities in the [0.47, 0.78] interval), while the rest of the spectra are covered by high-confidence predictions (Fig. 4i).

DreaMS Atlas

In this section, we provide a description of how the DreaMS Atlas is constructed. We start by outlining the process of selecting and annotating nodes for the DreaMS Atlas, followed by the process of connecting the nodes to form a graph structure.

The construction process begins with generating DreaMS embeddings for 76 million spectra comprising GeMS-C1 subset of the GeMS dataset. This subset represents LSH cluster representatives of 201 million GeMS-C spectra, covering the entire MassIVE GNPS repository. Spectra from blank samples, identified by specific suffixes in their names (e.g., "blank", "no_inj", "noinj", "empty", "solvent", or "wash"), are excluded. Additionally, we enrich individual nodes with DreaMS molecular property and fluorine presence predictions, along with relevant metadata obtained from the MassIVE repository, such as information about the study species, respective study description, and the instrument used for spectrum acquisition. Finally, we include embeddings of mass spectra from the MoNA and NIST20 spectral libraries. To avoid redundancy in the spectral libraries with respect to molecular structures, we merge spectra sharing identical canonical SMILES but differing in adduct species from both MoNA and NIST20, resulting in 79 thousand merged spectra from 819 thousand library entries.

Next, we employ the NN-Descent algorithm [65] to compute an approximate five-nearest-neighbor (5-NN) graph, where nodes represent DreaMS embeddings and edges represent similarities between these embeddings. To further refine the LSH clustering, 5-NN neighborhoods sharing DreaMS similarities above 0.9 are clustered into single nodes, and the k-NN graph is reconstructed for 34 million nodes representing the clusters. More precisely, to cluster the nodes, we iterate over all nodes sorted in descending order by their degrees and run a breadth-first search (BFS) from each node. The BFS stops if either an edge has a DreaMS similarity smaller than 0.9 or the DreaMS similarity between the starting node and the new candidate node is smaller than 0.9. All the nodes aggregated through the BFS are collapsed to a single cluster and are represented by a starting node. This algorithm allows us to cluster the graph in linear time. It is worth noting that by defining neighborhoods based on similarity thresholds rather than the number of hops, this algorithm adjusts the graph topology preventing over-representation of certain spectra.

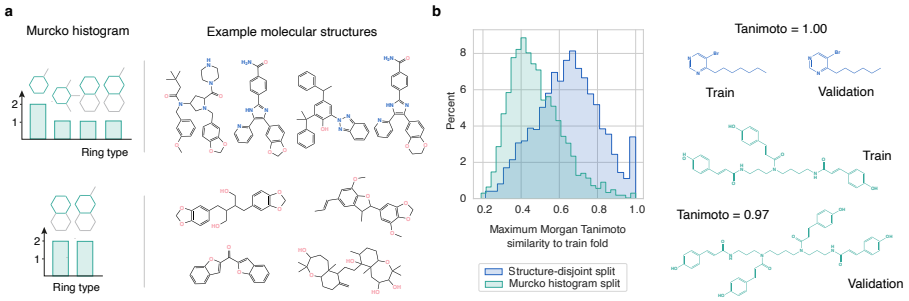
This procedure results in the creation of a final 5-NN graph representing the DreaMS Atlas. We utilize the PyNNDescent implementation of NN-Descent by McInnes et al. (https://github.com/lmcinnes/pynndescent), which provides functionalities for managing the vector database of the k-NN graph, such as querying the graph with new DreaMS embeddings not present in the DreaMS Atlas or extending the DreaMS Atlas with new embeddings.

Hyperparameters, ablation studies, implementation details

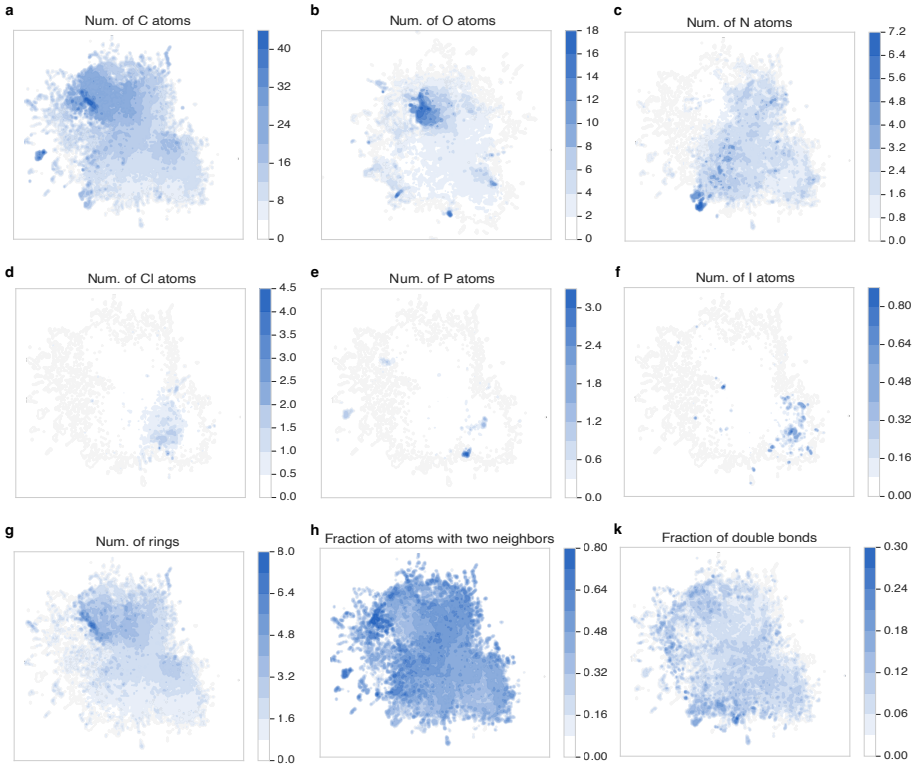
We report the hyperparameters used for pre-training and fine-tuning in Extended Data Table 2 and Extended Data Table 3, respectively. Extended Data Fig. 4 summarizes the key ablation studies highlighting three crucial features of our method: pre-training on the high-quality GeMS-A10 dataset, mass-tolerance Fourier features, and a masked m/z objective formulated as classification rather than regression. For both pre-training and fine-tuning, we used the Adam optimizer with default parameters [85].

All models were trained using either 4x AMD MI250X GPUs or 8x NVIDIA A100 GPUs in a distributed data parallel (DDP) mode. The final DreaMS model was pre-trained for 48 hours, while its fine-tuning runtime never exceeded several hours. With 8 NVIDIA A100 GPUs, the generation speed of embeddings (i.e., forward pass through the trained model) averages approximately 1.2 ± 0.002 million embeddings per hour, where the standard deviation is calculated based on twelve chunks comprising 79 unique mass spectra from GeMS-C.

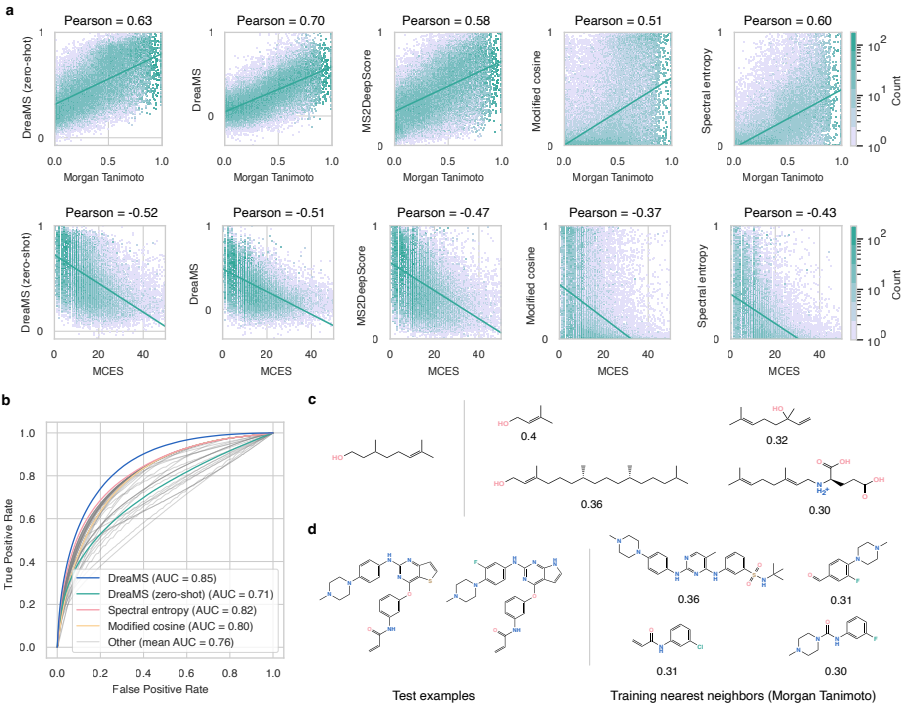
We used matchms [86] and pyOpenMS [87] Python libraries for processing mass spectra. All neural networks were implemented in PyTorch [88] and trained using PyTorch Lightning [89].



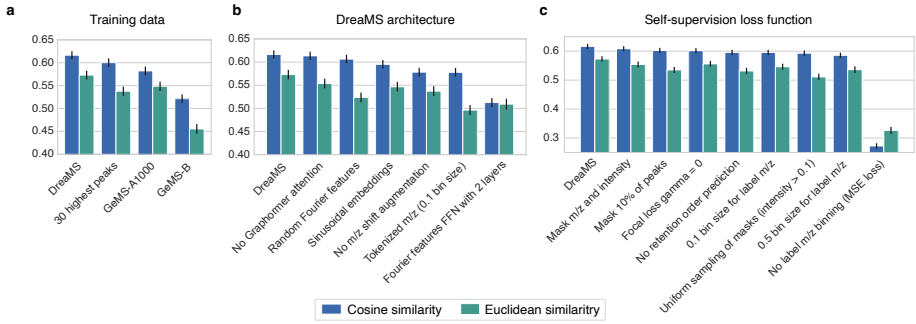
Extended Data Fig. 1 Our Murcko histograms splitting of spectral libraries surpasses the structure-disjoint approach. a, Schematic illustration of the Murcko histogram molecular representations. Different molecular structures shown on the right side have identical Murcko histograms shown on the left side. The domain of a Murcko histogram is defined in terms of rings (green structures) of different types based on the number of neighboring rings and linkers (grey structures). The corresponding values are the counts of such ring types. **b,** Evaluation of data leakage on structure-disjoint and Murcko histogram-disjoint splits of MoNA in terms of the maximum Morgan Tanimoto similarity to each training example computed for each validation example. Structure-disjoint splitting results in many leaking near-duplicate examples present in the validation set (maximum Tanimoto similarity > 0.95) whereas Murcko histograms-based splitting eliminates almost all such examples. The blue pair of structures represents such leaks, and the green pair demonstrates that the high-similarity inter-fold examples resulting from Murcko histogram splitting are rather a result of the imperfectness of Tanimoto similarity than the splitting algorithm. The mean of the Murcko histogram-disjoint similarity distribution is approximately one-half the value of the mean of the structure-disjoint-related distribution, implying that our approach to data splitting is better suited for evaluating model generalization.



Extended Data Fig. 2 Organization of DreaMS embeddings with respect to various structural properties of molecules. a-f, UMAP projections of DreaMS embeddings for 100,000 spectra from NIST20 visualized as contour plots colored by the number of atoms of different chemical elements. **g**–**k,** Identical projections as in **a**–**f,** but colored with respect to the topology of molecules. The number of rings and the fraction of atoms with two neighbors characterize the linearity and non-linearity of a molecule. Unlike other depicted properties, the fraction of double bonds is a property depending purely on the connectivity between atoms. Importantly, we show fractions to keep the coloring insensitive to the sizes of molecules, which is typically highly correlated with precursor m/z values of the mass spectra.



Extended Data Fig. 3 Extended benchmarking results. a, Cosine similarity in the space of DreaMS embeddings outperforms the supervised method MS2DeepScore and classic modified cosine and spectral entropy methods, as measured by two metrics: correlation to Morgan Tanimoto similarity and correlation to the maximum common edge subgraph distance [68]. **b,** Fine-tuned DreaMS demonstrates superior performance compared to standard spectral similarity in retrieving molecules from spectral libraries, considering a pool of candidates within a 10 ppm mass difference from the query precursor m/z. Importantly, although zero-shot cosine similarity on DreaMS excels in the correlation metrics presented in **a**, it lacks sensitivity to small structural differences among molecules. This observation motivates us to perform contrastive fine-tuning on challenging examples of molecules with similar masses. **c,d** The precursor molecules from the examples presented in Fig. 4c and Fig. 4j respectively, along with the four most similar training precursor molecules, as measured by Morgan Tanimoto similarity.



Extended Data Fig. 4 Ablation study of self-supervised pre-training. The evaluated metrics are Pearson correlation coefficients between cosine similarity and Euclidean similarity (the inverse of Euclidean distance) on DreaMS embeddings with Morgan Tanimoto molecular similarity. The bars display maximum values of the metrics achieved through the course of pre-training with each of the model configurations. The underlying dataset comprises 5,000 pairs of spectra from NIST20 sampled to maximize the entropy of the Tanimoto similarity distribution. The error bars represent standard deviations within 1,000 bootstrap samples. DreaMS represents the final model presented in this work, trained on the GeMS-A10 dataset using the sixty highest peaks, Graphormer-like attention mechanism, Fourier features (5 layers for subsequent feed-forward network), m/z shift augmentations, masking 30% of m/z values (not masking intensities) sampled proportionally to corresponding intensities, focal loss with gamma equal to 5, retention order prediction, and a 0.05 bin size for m/z labels. Sinusoidal embeddings and tokenized m/z refer to peak representations proposed by Voronov et al. [14, 25].

Method	Top 1	Top 5	Top 10	Top 20	Top 50	Top 100	Top 200	Cos. similarity
FFN fingerprint	17.309	37.121	45.611	54.634	63.946	71.329	77.359	0.537
FFN contrastive	20.632	44.996	54.389	63.536	75.062	80.558	86.095	-
MIST fingerprint	29.368	55.332	63.536	72.231	80.476	85.726	89.418	0.695
MIST contrastive	28.384	55.373	65.217	72.970	81.255	85.480	89.377	-
MIST contrastive + fingerprint	30.703	58.120	68.927	75.709	84.094	87.916	92.355	-
DreaMS fingerprint (ours)	32.731	59.352	67.719	75.390	82.404	87.121	90.771	0.646

Extended Data Table. 1 Full test metrics on MIST PubChem retrieval benchmark. The "Top k" columns stand for the retrieval accuracy@k metrics reported in percents (higher is better). "Cos. similarity" shows the cosine similarity (higher is better) between predicted and ground-truth fingerprints (i.e., inverted test loss). Methods with the "fingerprint" suffix denote the methods directly predicting molecular fingerprints whereas "contrastive" means that the training procedure involves batches of PubChem molecules with the same molecular formula and learns to correctly rank candidates via a noise contrastive estimation (NCE) loss function [31, 90]. We do not experiment with the contrastive extension for our model since it outperforms contrastive methods on top 1 and top 5 metrics without considering additional PubChem molecules. Interestingly, DreaMS underperforms MIST in terms of the cosine similarity despite performing better in retrieval. A similar performance on the same benchmark is observed with SIRIUS, which outperforms MIST in retrieval despite being less accurate in fingerprint prediction [31].

Hyperparameter	**Values**
Learning rate	$5 \cdot 10^{-5}, 9 \cdot 10^{-5}, \mathbf{1 \cdot 10^{-4}}, 2 \cdot 10^{-4}, 3 \cdot 10^{-4}$
Number of warmup steps [53]	0, **5000**, 20000
Batch size	1024, 2048, **4096**
Number of transformer layers l	1, 5, **7**, 11
Number of attention heads	4, **8**, 12, 16
Transformer hidden dimensionality d	512, 768, **1024**
Fourier features dimensionality d_m	24, 512, **980**
Peak dimensionality d_p	**24**, 512, 980
FFN_F depth	2, 4, **5**
FFN_F hidden dimensionality	256, **512**
FFN_P depth	**1**, 2, 3
Attention mechanism	dot-product, additive [91], **Graphormer**
Dropout	0.0, **0.1**, 0.5
Weight decay	**0.0**, $1 \cdot 10^{-5}$
Fraction of masked peaks	0.1, 0.2, **0.3**, 0.4, 0.5
Mask sampling strategy	uniform (intensity > 10%), **intensity proportional**
Deterministic mask sampling	**True**, False
Retention order loss weight	0.0, **0.2**, 0.5
Focal loss γ	0, 0.5, 2, **5**
Dataset	**GeMS-A10**, GeMS-A1000, GeMS-A, GeMS-B
Training float precision	**32 bits**, 64 bits

Extended Data Table. 2 Explored pre-training hyperparameters. The optimal values, used for the extraction of embeddings, zero-shot predictions, and further fine-tuning, are highlighted in bold.

Spectral similarity

Hyperparameter	**Values**
Learning rate	$3 \cdot 10^{-6}$, $\mathbf{5 \cdot 10^{-6}}$, $1 \cdot 10^{-5}$
Batch size	**32**, 64
Head type	**Linear**
Triplet margin Δ	0.05, **0.1**, 0.2, 0.5

Molecular fingerprint prediction

Hyperparameter	**Values**
Learning rate	$2 \cdot 10^{-5}$, $3 \cdot 10^{-5}$, $\mathbf{4 \cdot 10^{-5}}$
Batch size	16, **32**, 64
Head type	Linear, **DeepSets**

Molecular property prediction

Hyperparameter	**Values**
Learning rate	$3 \cdot 10^{-4}$, $\mathbf{3 \cdot 10^{-5}}$
Batch size	128, **512**
Head type	**Linear**

Fluorine detection

Hyperparameter	**Values**
Learning rate	$3 \cdot 10^{-5}$, $\mathbf{5 \cdot 10^{-5}}$
Batch size	64, **128**
Head type	**Linear**
Focal loss α	0.5, 0.6, **0.8**
Focal loss γ	**0.5**, 1, 2

Extended Data Table. 3 Explored fine-tuning hyperparameters. The optimal values, used for test predictions, are highlighted in bold.

MS^2 data	**Data type**
M/z values	float64
Intensities	float32
MS level	int8
RT	float32
Charge	int8
Polarity	int8
Precursor m/z	float32
Window lbound	float32
Window ubound	float32
CID energy	float32
Spectrum type	int8
Ion injection time	float32
Definition string	utf-8 str
Precursor id	int32

Metadata	**Data type**
File name	utf-8 str
Instrument name	utf-8 str
MS level order	utf-8 str
$\|X_1\|$	int64
$\|X_2\|$	int64
$\textsc{Median}(A)$	float64

Precursor data	**Data type**
M/z values	float64
Intensities	float32
RT	float32
Ion injection time	float32
Id	int32

Extended Data Table. 4 Specification of the GeMS `.hdf5` data format. "MS^2 data" and "Precursor data" are `.hdf5` groups whereas "Metadata" entities are `.hdf5` attributes. All tensors are one-dimensional of the length equal to the number of collected spectra. The only exception is "M/z values" and "Intensities" which are two-dimensional arrays of the number of spectra by the number of peaks shape. We retain 128 highest peaks and pad the array with zeros. $|X_1|$, $|X_2|$, and $\textsc{Median}(A)$ correspond to the intermediate values and the output of Algorithm 1. "Window lbound" and "Window ubound" correspond to the lower and upper bounds of the MS^1 isolation window. "Definition string" is a spectrum metadata summary string available in the data from Thermo instrument.