MassSpecGym: A benchmark for the discovery and identification of molecules

38th Conference on Neural Information Processing Systems (NeurIPS 2024)

Bushuiev, Roman; Bushuiev, Anton; de Jonge, Niek F.; Young, Adamo; Kretschmer, Fleming et al

https://proceedings.neurips.cc/paper_files/paper/2024/hash/c6c31413d5c53b7d1c343c1498734b0f-Abstract-Datasets_and_Benchmarks_Track.html

MassSpecGym: A benchmark for the discovery and identification of molecules

Roman Bushuiev[1,2], **Anton Bushuiev**[2], **Niek F. de Jonge**[3], **Adamo Young**[4], **Fleming Kretschmer**[5], **Raman Samusevich**[1,2], **Janne Heirman**[6], **Fei Wang**[7,8], **Luke Zhang**[9], **Kai Dührkop**[5], **Marcus Ludwig**[10], **Nils A. Haupt**[5], **Apurva Kalia**[11], **Corinna Brungs**[1], **Robin Schmid**[1], **Russell Greiner**[7,8], **Bo Wang**[4], **David S. Wishart**[7,12], **Li-Ping Liu**[11], **Juho Rousu**[13], **Wout Bittremieux**[6], **Hannes Rost**[9], **Tytus D. Mak**[14], **Soha Hassoun**[11,15], **Florian Huber**[16], **Justin J.J. van der Hooft**[3,17], **Michael A. Stravs**[18], **Sebastian Böcker**[5], **Josef Sivic**[2], **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, [3]Bioinformatics Group, Wageningen University & Research, [4]Department of Computer Science, University of Toronto, [5]Chair for Bioinformatics, Institute for Computer Science, Friedrich Schiller University Jena, [6]Department of Computer Science, University of Antwerp, [7]Department of computing science, University of Alberta, [8]Alberta Machine Intelligence Institute, [9]Department of Molecular Genetics, University of Toronto, [10]Bright Giant GmbH, [11]Department of Computer Science, Tufts University, [12]Department of Biological Sciences, University of Alberta, [13]Department of Computer Science, Aalto University, [14]Mass Spectrometry Data Center, National Institute of Standards and Technology, [15]Department of Chemical and Biological Engineering, Tufts University, [16]Centre for Digitalisation and Digitality, University of Applied Sciences Düsseldorf, [17]Department of Biochemistry, University of Johannesburg, [18]Eawag: Swiss Federal Institute of Aquatic Science and Technology

Abstract

The discovery and identification of molecules in biological and environmental samples is crucial for advancing biomedical and chemical sciences. Tandem mass spectrometry (MS/MS) is the leading technique for high-throughput elucidation of molecular structures. However, decoding a molecular structure from its mass spectrum is exceptionally challenging, even when performed by human experts. As a result, the vast majority of acquired MS/MS spectra remain uninterpreted, thereby limiting our understanding of the underlying (bio)chemical processes. Despite decades of progress in machine learning applications for predicting molecular structures from MS/MS spectra, the development of new methods is severely hindered by the lack of standard datasets and evaluation protocols. To address this problem, we propose MassSpecGym – the first comprehensive benchmark for the discovery and identification of molecules from MS/MS data. Our benchmark comprises the largest publicly available collection of high-quality labeled MS/MS spectra and defines three MS/MS annotation challenges: *de novo* molecular structure generation, molecule retrieval, and spectrum simulation. It includes new evaluation metrics and a generalization-demanding data split, therefore standardizing the MS/MS annotation tasks and rendering the problem accessible to the broad machine learning community. MassSpecGym is publicly available at `https://github.com/pluskal-lab/MassSpecGym`.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

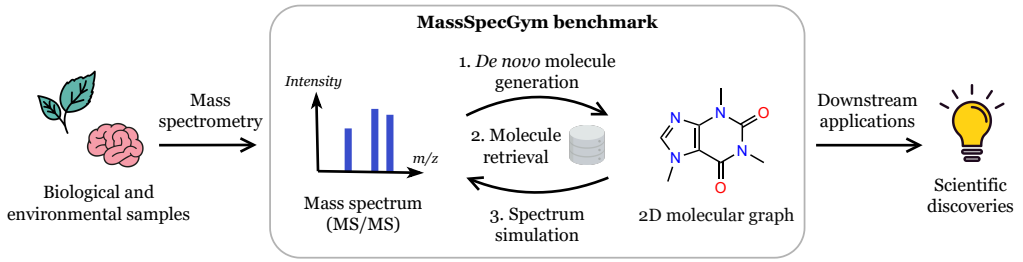

Figure 1: **MassSpecGym provides three challenges for benchmarking the discovery and identification of new molecules from MS/MS spectra**. The provided challenges abstract the process of scientific discovery from biological and environmental samples into well-defined machine learning problems.

1 Introduction

The discovery and identification of small molecules profoundly influence numerous scientific fields, including organic chemistry [1], molecular biology [2], drug development [3], disease diagnosis [4], environmental analysis [5], and space exploration [6]. Despite significant progress, it is estimated that only a small fraction of molecules across the kingdoms of life have been discovered [7]. Tandem mass spectrometry (MS/MS) is the most widely used technique for elucidating molecular structures from biological and environmental samples, supporting a wide range of applications in biotechnology and medicine [8]. In drug development, MS/MS is crucial for identifying novel bioactive compounds [9], such as those targeting cancer and infectious diseases [7]. MS/MS also plays a key role in clinical settings for determining appropriate drug dosages and assessing potential side effects [10]. In environmental analysis, it enables the detection of pollutants at trace levels, which is vital for monitoring and preserving environmental health [11]. Moreover, MS/MS addresses various challenges in structural biology, including the discovery of ligands that bind to target proteins [12] and the elucidation of metabolic pathways [13].

When analyzing a sample, a mass spectrometer typically generates thousands of tandem mass spectra, each characterizing a specific molecule present in the sample. While the annotation of mass spectra with molecular structures is inherent to mass spectrometry, it remains a significant challenge. From typical samples of interest, typically less than 10% of MS/MS spectra are annotated using state-of-the-art methods [14, 15]. As a result, the natural chemical space remains largely unexplored, thereby hindering scientific advancements.

To generate an MS/MS spectrum, a mass spectrometer follows an intricate multi-step procedure. First, the instrument ionizes the molecule using methods such as electrospray ionization (ESI). During this process, the molecule gains additional atoms, known as the ionization adduct. Subsequently, the ionized molecule (often referred to as precursor ion) is fragmented using collision-induced dissociation (CID), higher energy collisional dissociation (HCD), or other fragmentation method [16]. Finally, for each individual fragment ion, the instrument records its (i) mass-to-charge ratio (m/z value; the charge is typically equal to one for small molecules) and (ii) its corresponding abundance (signal intensity). The collection of these two-dimensional data points, characterizing the molecule as a distribution of fragment masses, is referred to as a tandem mass spectrum, MS/MS spectrum, or MS^2 spectrum.

The most notable progress in MS/MS annotation has been achieved by machine learning methods augmented with combinatorial optimization and domain expertise [17, 18]. However, these methods have not seen significant improvements in recent years due to their lack of scalability and small return of increased human knowledge. In contrast, recent years have witnessed numerous purely data-driven deep learning models performing competitively or even surpassing the classic approaches [19, 20, 21, 22, 23, 24, 25, 26, 27]. Nevertheless, the development of this new generation of modern machine learning methods for MS/MS spectrum annotation is currently hindered by multiple factors. These factors include the heterogeneity of data acquired under different mass spectrometry settings, the scarcity of high-quality annotated spectra, variations in data pre-processing techniques, inconsistencies in data splitting methods resulting in data leakage, differences in approaches to MS/MS annotation, varying evaluation metrics, and the proprietary nature of many datasets. As a result, developing a machine learning algorithm for mass spectrum annotation currently necessitates

mass spectrometry domain expertise, rigorous data preparation, and the reevaluation of existing methods for benchmarking purposes.

At the same time, dataset collection and benchmarking efforts have been one of the key drivers responsible for breakthrough progress in machine-learning-driven fields, for example: ImageNet [28], SQuAD [29], Gym [30], ProteinGym [31, 32], and OGB [33]. Inspired by these efforts, we propose MassSpecGym – a new public dataset of MS/MS spectra and a unified benchmarking protocol for MS/MS spectrum annotation (Figure 1). Our dataset provides a standardized collection of 231 thousand high-quality mass spectra representing 29 thousand unique molecular structures, making it the largest publicly available dataset. 10 thousand molecules (33%) present in the dataset are derived from our newly measured in-house data (i.e., MSnLib library presented in [34]). Additionally, we provide a curated selection of large unlabeled datasets of mass spectra and molecules allowing for the combination of supervised and unsupervised methods [20, 35]. Importantly, we develop a new splitting procedure based on the edit distance of molecular structures and divide our dataset into non-leaking train-validation-test folds. The MassSpecGym benchmark defines three MS/MS annotation challenges: *de novo* molecular structure generation, molecule retrieval, and spectrum simulation. We make each of the challenges easily accessible to the broad machine learning community by providing MassSpecGym through a user-friendly interface leveraging PyTorch Lightning and Hugging Face platforms[1]. Users can build new models on top of the prepared components and submit their results to the *Papers With Code* leaderboard. We anticipate that our unified benchmark will have a significant impact on the community by enabling reproducible research and accelerating the development of new MS/MS spectrum annotation methods.

2 Related work

Labeled MS/MS data. The creation of spectral libraries is driven by the desire to facilitate the annotation of a measured query spectrum [39, 40]. A spectral library catalogues a molecule and one or more of its spectra that are measured under different mass spectrometry instrument conditions. There are in-house private spectral libraries, commercial libraries, and openly accessible crowd-sourced libraries. MassBank [38], MassBank of North America (MoNA) [37] and GNPS [36] are the three largest crowd-sourced libraries comprising tens of thousands of molecules in total. The National Institute of Standards and Technology (NIST) provides a variety of for-purchase spectral libraries comprising up to 52 thousand compounds. However, NIST libraries are not available for machine learning due to licensing restrictions. A similar situation exists with mzCloud [41], which provides MS/MS spectra for 32 thousand compounds but cannot be downloaded and used outside its native web interface.

Table 1: **MassSpecGym is the largest publicly available dataset of high-quality labeled MS/MS spectra.** Our quality assessment workflow eliminates noisy or corrupted spectra and ensures reliable molecular labels and metadata (Section 3.3). The "Split" column highlights that, unlike other large-scale datasets, MassSpecGym provides a pre-defined data split.

Dataset	Spectra	High-quality spectra	Molecules	Split
GNPS [36]	**322K**	104K	16K	✗
MoNA [37]	98K	62K	10K	✗
MassBank [38]	62K	58K	4K	✗
MIST CANOPUS [19]	11K	$\leq$ 11K	$\leq$ 9K	✓
MassSpecGym (ours)	231K	**231K**	**29K**	✓

The availability of spectral libraries has provided labeled datasets for supervised machine learning, but there are many challenges. These libraries are relatively small, covering only thousands to tens of thousands of molecules. Consequently, many annotation tools combine data from various libraries, including proprietary sources, limiting reproducibility and introducing biases. Public crowd-sourced datasets often contain low-quality, noisy mass spectra or invalid metadata, necessitating custom pre-processing and filtering techniques. While these techniques aim to improve dataset quality, they often limit the applicability and reproducibility of the corresponding machine learning methods. Additionally, the heterogeneity and non-standardization of mass spectrometry instruments and parameters challenge effective learning from spectral libraries. Our MassSpecGym dataset offers the first carefully curated and standardized collection of MS/MS spectra, maintaining high quality and surpassing existing datasets in size (Table 1).

[1] `https://github.com/pluskal-lab/MassSpecGym`

Train-validation-test splitting of MS/MS data. Most of the previous studies split labeled MS/MS data such that molecules with identical planar structures do not appear in different training, validation, and test sets [17, 35, 24, 19, 21]. This is achieved by using distinct 2D InChiKey hash descriptions of molecules for each data fold. However, this method can be compromised by minor structural modifications often found in spectral libraries as a result of, for example, click chemistry [42]. Our MassSpecGym benchmark has undergone extensive vetting in terms of data splitting. In this work, to prevent data leakage and to accurately assess model generalization to novel molecules, we develop a data splitting strategy that guarantees that there are no leaks with the chemical bond edit distance (i.e., MCES distance [43]) less than 10 (Figure 2).

Benchmarking MS/MS annotation. Currently, there are no comprehensive and standardized datasets available for the development and evaluation of models predicting spectra or molecular structures. One recently utilized dataset for benchmarking is MIST CANOPUS [19, 35], which was curated to ensure an even distribution of chemical classes [44]. However, this dataset is relatively small, comprising only 9 thousand molecules and 11 thousand spectra, and employs a data split based on 2D InChIKey, a method resulting in data leakage (Figure 2).

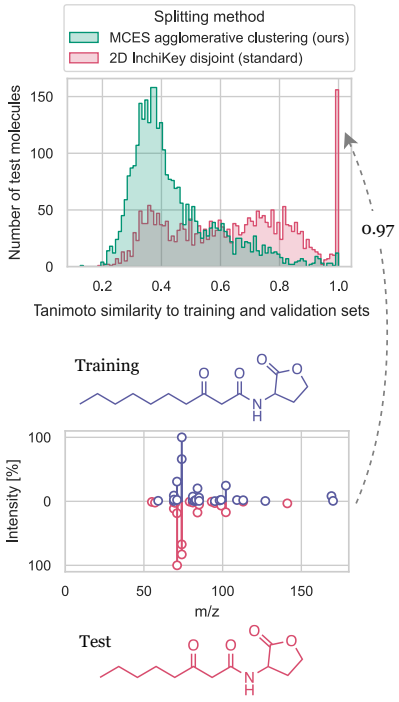

Figure 2: **Our MCES-based data split resolves the data-leakage issue abundant in prior work**. The standard approach separates molecules with identical planar structures (2D InChIKeys) into different folds, disregarding minor molecular modifications. This leads to near-duplicate test molecules (with Tanimoto similarity > 0.85) being leaked from the training data (red), as shown in the example below. In contrast, our approach maximizes molecular edit distance (MCES) between training and test sets, ensuring distinct data folds (green).

The Critical Assessment of Small Molecule Identification (CASMI) series [45] is another example of a recent benchmarking initiative. However, the CASMI challenge is held only once every two years at best, limiting opportunities for continuous evaluation and benchmarking. Additionally, the CASMI datasets are relatively small, comprising several hundred spectra representing challenging test cases. Participation in the challenge also demands significant mass spectrometry domain expertise for preprocessing the data into the format suitable for machine learning. In contrast, our proposed MassSpecGym is based on the new largest publicly available dataset (Table 1) and is designed to be machine learning-ready, thereby addressing the limitations inherent in the MIST CANOPUS and CASMI benchmarks.

3 MassSpecGym benchmark

This section describes the construction of the MassSpecGym benchmark. First, we define three challenges of mass spectrum annotation along with the corresponding evaluation metrics (Section 3.2). Then, we describe the collection and processing of the underlying dataset of mass spectra and analyze its composition (Section 3.3). Finally, we outline our procedure for splitting the dataset into train-validation-test folds and demonstrate its generalization-demanding nature (Section 3.4). Please see the details on the construction of MassSpecGym in Supplementary Information.

3.1 Motivation for the challenges

***De novo* molecule generation.** The first challenge is the *de novo* prediction of a molecular graph from an MS/MS mass spectrum. This challenge can be compared to the goal of AlphaFold [46] but instead of predicting protein structures from their sequences, the task here is to predict small molecule structures from their MS/MS spectra. As such, this task represents a grand challenge in computational mass spectrometry, given its potential to drive the discovery of novel natural products, drug metabolites, environmental transformation products, and other crucial molecules [14]. A model

that can accurately predict molecular structures from MS/MS spectra could significantly advance our understanding of biology by enabling the annotation of metabolomes of uncharacterized organisms [7].

Molecule retrieval. The second challenge focuses on retrieving a molecular graph from a molecular database given a mass spectrum, rather than generating a completely new molecule. This scenario is common in practice when determining if a sample contains specific compounds, such as pesticides, environmental pollutants, or other known substances [11]. This approach is also relevant in drug design, particularly in affinity selection–mass spectrometry, where protein binders are identified from combinatorial libraries of ligands [12].

Spectrum simulation. The third challenge, called spectrum simulation, is the inverse problem of predicting a mass spectrum from a molecular graph. This task has two main motivations. First, it enhances the understanding of MS/MS fragmentation mechanisms in organic chemistry, leading to more precise predictions of how molecules will behave under various conditions. This insight can aid in the design of novel compounds and the optimization of synthetic pathways [47]. Second, it allows for pseudolabeling, expanding training datasets for machine learning models by generating synthetic spectra, which can improve model performance when experimental data is limited [48].

3.2 Definition of the challenges

***De novo* molecule generation.** The task of *de novo* molecular generation is to generate a molecular structure from a mass spectrum. Formally, the input is a mass spectrum $X \subset \mathbb{R}_+ \times (0, 1]$, consisting of a set of two-dimensional points (referred to as signals or peaks) representing m/z (mass-to-charge) values and their corresponding intensities, which are normalized by dividing each by the maximum intensity. Intuitively, these points describe the abundance of molecular fragments with different masses. The goal is to generate a molecular graph $\hat{G} = (V, E)$, where vertices $V \in \mathbb{V}^N$, $|\mathbb{V}| = 118$ is a set of N atoms from the vocabulary of 118 chemical elements (or, for example, 10 most common ones [24]) characterized by the periodic table, and $E \in \mathbb{E}^M$, $|\mathbb{E}| = 4$ is a set of M edges from the vocabulary of 4 chemical bonds between atoms: single, double, triple, and aromatic [33]. Note that we do not model the 3D coordinates of chemical graphs, as the information in MS/MS spectra is typically insufficient for predicting exact molecular conformations [49].

Given the complexity of *de novo* generation, we propose an additional, simpler, challenge where chemical formulae are provided as input, meaning the set of vertices V is known. In practice, chemical formulae can be derived with high accuracy by utilizing MS^1 mass spectra, an orthogonal data source to MS/MS [50, 51]. Since working with MS^1 data is typically based on combinatorial optimization rather than machine learning [52], our benchmark directly provides chemical formulae instead of MS^1 spectra, imitating the output of the MS^1 spectra processing pipelines. However, we present this scenario as a bonus challenge because chemical formula prediction remains a partially unsolved problem. For example, elements such as fluorine, which have only one stable isotope, cannot be derived from MS^1 data alone and still pose challenges with MS/MS data [20].

While each mass spectrum is a measurement on a specific compound, the spectrum may not contain all the necessary information to fully reconstruct the molecular structure as the spectrum is a partial view of the measured compound. Therefore, our approach acknowledges this complexity and permits multiple plausible molecular structures corresponding to a given spectrum. To this end, we formulate the problem as predicting a set of k graphs $\hat{\mathcal{G}}_k = \{\hat{G}_1, \ldots, \hat{G}_k\}$ rather than a single solution $\hat{G}$. These graphs can be sampled randomly from a model or selected as the top-k predictions from a larger set, if a scoring function is available. This approach reflects the inherent uncertainty and challenges of accurately predicting the correct molecular graph from spectral data.

We evaluate the correspondence between the generated molecular graphs $\hat{\mathcal{G}}_k$ and the ground-truth graph G using three metrics. Ideally, the set of predictions includes the ground-truth graph, which we assess by measuring

$$\text{Top-}k \text{ accuracy: } \mathbb{1}\{G \in \hat{\mathcal{G}}_k\}, \tag{1}$$

averaged across all test examples. In the equation, $\mathbb{1}$ is the indicator function returning 1 if the condition is true and 0 otherwise. The top-k accuracy varies between 0 and 1, where 0 corresponds to none of the test samples having the ground truth graph among the top-k prediction and 1 corresponds

to all test samples having the ground truth graph among the top-k predictions. Given the difficulty of predicting the exact graph, we also assess the similarity between predicted molecules and the true molecule using two molecular similarity measures. First, we use the maximum common edge subgraph (MCES) metric [43], which is an edit distance on molecular graphs. Specifically, we evaluate how close the most similar prediction is to the true molecule in terms of the MCES distance across top-k predictions (we evaluate $k \in \{1, 10\}$):

$$\text{Top-}k \text{ MCES: } \min_{\hat{G} \in \hat{\mathcal{G}}_k} MCES(\hat{G}, G), \tag{2}$$

averaged across test examples. The MCES distance is 0 when two graphs are identical, and increasing values correspond to decreasing similarity. We also use the Tanimoto similarity (or Jaccard coefficient) on the Morgan fingerprints of molecules [53], which measures how well a generative model recognizes true molecular fragments:

$$\text{Top-}k \text{ Tanimoto: } \max_{\hat{G} \in \hat{\mathcal{G}}_k} Tanimoto(\hat{G}, G). \tag{3}$$

The Tanimoto similarity between two molecules ranges from 0 to 1, where a value of 1 indicates identical molecules.

Molecule retrieval. In practice, *de novo* molecule generation is often infeasible due to the combinatorial complexity of the solution space. An alternative and practically relevant problem is molecule retrieval, which is to rank candidate molecular graphs (from a chemical database) for a given input spectrum. Formally, given a mass spectrum $X \subset \mathbb{R}_+ \times (0, 1]$, the task is to order a set of candidate graphs $\mathcal{C} = \{G_1, \ldots, G_n\}$ such that the correct molecular graph $G \in \mathcal{C}$ has the lowest index.

Chemical databases may contain millions of molecules, e.g., the PubChem database has over 118 million molecules [54], making it impractical to sort the entire set. However, since the mass of the true molecule can be derived from an MS/MS spectrum, the candidate set $\mathcal{C}$ can be constructed to include only molecules with same masses as the true one (within an acceptable experimental error range). To standardize the task across examples, we limit $|C| \leq 256$ candidates per spectrum, sampled randomly if more molecules with the same mass are available. Additionally, similar to the *de novo* generation task, we define a bonus challenge where the set V is known via the molecular formula, allowing further pruning of $\mathcal{C}$ to include only graphs with the given nodes V.

We evaluate molecule retrieval using standard information retrieval metrics, as well as the molecular similarity of the top hit with the true molecule. Specifically, we measure:

$$\text{Hit rate @ } k\text{: } \mathbb{1}\{G \in \mathcal{C}_k\}, \tag{4}$$

averaged across all test examples, where $\mathcal{C}_k \subset \mathcal{C}$ is the set of top-k hits sorted by the model and $\mathbb{1}$ is the indicator function. The hit rate @ k ranges from 0 to 1, with 1 indicating perfect performance, meaning all true molecules were correctly retrieved among the top k candidates. Additionally, we evaluate the average similarity of the top-1 hit $G_1 \in \mathcal{C}$ with the ground truth molecule G by measuring the maximum common edge subgraph (MCES) distance [43]:

$$\text{MCES @ 1: } MCES(G_1, G). \tag{5}$$

The MCES @ 1 value is 0 if the top-1 retrieved candidate is exactly the true molecule, with higher positive values indicating lower similarity between the molecules.

Spectrum simulation. In contrast to the above spectrum-to-molecule tasks, spectrum simulation is the inverse problem of predicting an MS/MS spectrum. The input is a molecular graph $G = (V, E)$ and the measurement parameters $I \in \{I_1, I_2, ..., I_n\}$, $A \in \{A_1, A_2, ..., A_m\}$, and $C \in \mathbb{R}_+$, where I represents the type of instrument used, A represents the adduct associated with the precursor ion, and C is the amount of energy used during fragmentation (the collision energy, measured in electronvolts or eV). The output is a predicted mass spectrum $\hat{X} \subset \mathbb{R}_+ \times (0, 1]$. To limit complexity from extra parameters tangential to the issue, we restrict the task to spectra with the most abundant adduct ($A =$ [M+H]$^+$) and simplify the instrument types to the two principal technologies ($I \in$ {QTOF, Orbitrap}).

Typically, $\hat{X}$ and the true spectrum X have binned representations $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^d$, where d is the number of bins. Instead of listing exact locations of m/z peaks and their intensities, they discretize the space

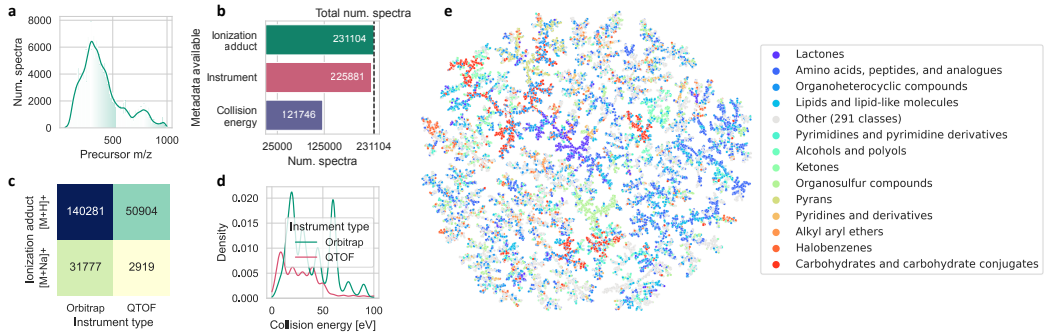

Figure 3: **MassSpecGym provides a diverse and highly-standardized dataset of MS/MS spectra.** The histogram of precursor m/z values (**a**) and a TMAP [58] projection of precursor molecules (**e**) demonstrate a rich coverage of molecular masses and chemical classes [44] in MassSpecGym. Unlike other spectral libraries, our dataset is highly standardized in terms of mass spectrometry metadata. Each spectrum has an associated ionization adduct, either [M+H]+ or [M+Na]+, and nearly all spectra (98%) are linked to MS instruments, either Orbitrap or QTOF (**b, c**). Approximately half of the dataset entries (53%) contain normalized collision energies (**b, d**).

into a series of m/z bins to store peaks in their approximate positions [55, 56, 57]. The selection of bin size, and by extension d, requires consideration: in this benchmark, we choose a bin size of 0.01 Da and a maximum m/z of 1005 Da, resulting in $d = 100500$. These values are precise enough to retain important information about peak accuracy without becoming overly sensitive to measurement error. Additionally, we exclude potential precursor signals from both ground truths and predictions in the benchmark since there is a tendency for strong precursor signals to inflate model performance.

An evaluation metric for the quality of the prediction is the cosine similarity between the predicted binned spectrum $\hat{\mathbf{x}}$ and the ground truth $\mathbf{x}$ (Equation 6), averaged across all graph-spectra pairs:

$$\text{Cosine similarity:} \frac{\hat{\mathbf{x}}^T \mathbf{x}}{\|\hat{\mathbf{x}}\| \|\mathbf{x}\|}. \tag{6}$$

Cosine similarity between two spectra ranges from 0 to 1, where 1 corresponds to a perfect prediction. Jensen-Shannon similarity is reported as an additional metric (see Supplementary Information).

A key application of accurately predicting spectra from molecules is in molecular retrieval [55, 21, 23]. Accurate and scalable models enable the automatic annotation of molecular databases, bolstering the coverage of existing spectral libraries. Therefore, we can additionally use an analagous setup as a downstream task to evaluate spectrum predictions. Similarly to the molecule retrieval task defined previously, for a molecule-spectrum pair (G, X), the set of candidates $\mathcal{C}$ comprises of G and the set of molecules from a chemical database most similar to G. For each molecule $G_i \in \mathcal{C}$, we predict a spectrum $\hat{\mathbf{x}}_i$ and rank all candidates by decreasing cosine similarity between $\hat{\mathbf{x}}_i$ and $\mathbf{x}$. We evaluate the model by the rate at which G is ranked in the top-k hits in the sorted $\mathcal{C}$, using the same hit rate @ k metrics as defined in the previous section.

3.3 Dataset collection

To construct the MassSpecGym dataset, we first exhaustively collected MS/MS spectra from the largest publicly available spectral libraries: MoNA [37], MassBank [38], and GNPS [36] (downloaded from the official websites on May 27, 2024), as well as from our in-house data [34]. We then deduplicated and cleaned the spectra by applying a series of matchms filtering criteria [59]. These criteria are mainly based on the protocol described in [60] and involve additional filters to better standardize the dataset, such as keeping only spectra of molecules with m/z < 1000 or spectra with fewer than 1000 signals. To ensure the high quality of the dataset, we applied additional criteria, such as removing all spectra where more than 50% of the total intensity cannot be explained by combinatorially decomposing molecular mass into plausible chemical subformulae [61]. We preprocessed the mass spectra by removing signals estimated to be instrument noise. Finally, we standardized both molecular structures and mass spectra, and harmonized metadata entries, inferring

missing or incorrect values where possible [62, 60]. Figure 3 shows that our resultant dataset is rich in terms of molecular structures and highly standardized in terms of mass spectrometry metadata.

Additionally, we provide curated unlabeled datasets of mass spectra and molecules. For the mass spectra, we provide the GeMS-A10 dataset, a deduplicated collection of 24 million high-quality mass spectra mined from the MassIVE repository [20]. For the molecules, we provide (i) 1 million molecules of biological and environmental origin, including collections of natural products, pesticides, industrial chemicals, food additives, and other compounds [43], (ii) 4 million molecules spanning a diverse range of chemical classes [35], and (iii) all 118 million molecules from PubChem [63] (downloaded from the official website on May 31, 2024).

First, we utilize these three molecular datasets to construct retrieval candidates $\mathcal{C}$ for the molecule retrieval and spectrum simulation tasks (Section 3.2). For each spectrum-molecule pair, we iteratively sample molecules with the same mass as the query molecule from (i), followed by (ii) and (iii) until the maximum number of candidates $|\mathcal{C}| = 256$ is reached. The sequence of the datasets used for sampling reflects the relevance of their composition for mass spectrometry applications. A similar procedure is applied for the bonus challenge, where candidates are selected based on identical molecular formulae.

Second, when developing new methods that leverage unlabeled data, we anticipate that users will rely solely on the following two datasets: the GeMS-A10 dataset of unlabeled MS/MS spectra and a refined subset of the unlabeled 4 million-molecule dataset. The molecular dataset has been refined by excluding any molecules with an MCES distance of less than two from any molecule in the test fold of MassSpecGym. This refinement is intended to reduce the potential for data leakage, particularly when used in the context of the *de novo* generation challenge.

3.4 Dataset splitting

We split our dataset using MCES distances between molecular graphs corresponding to mass spectra. Specifically, we group all 29 thousand unique molecules into training, validation, and test folds using agglomerative clustering with MCES as the metric and the minimum distance as the linkage criterion. By setting the linkage distance threshold to 10, our approach ensures that no molecules have an edge edit distance of less than 10 between different data folds. Figure 2 shows that our method significantly surpasses the commonly applied 2D InChIKey disjoint approach, used in nearly all related works, in terms of preventing data leakage. Additionally, we stratify the spectra by instrument types, collision energies, ionization adducts, and the frequency of the molecules in the entire dataset, resulting in balanced folds with respect to metadata. A more detailed description of the splitting and additional analysis is available in the Supplementary Information.

4 Experiments

4.1 Baseline models

To establish reference performance across the tasks, we evaluate an initial set of representative baseline methods summarized in this section. Please see Supplementary Information for details.

***De novo* molecule generation.** For the *de novo* molecule generation challenge, we begin by implementing a baseline model based on prior chemical knowledge, referred to as **Random chemical generation**, which produces random chemically valid molecules given specific molecular masses or formulae. This baseline uses combinatorial and graph theory algorithms, drawing from statistics derived from the training data. To complement this domain-knowledge baseline, we also implement two Transformer models [64]. These models encode two-dimensional continuous tokens, representing m/z–intensity value pairs of MS/MS spectra, and decode string representations of molecular graphs. The first model, **SMILES Transformer**, decodes molecules as byte-pair-encoded [65] SMILES strings [66]. The second model, **SELFIES Transformer**, decodes molecules as SELFIES strings [67], offering the advantage of always producing valid chemical structures. We do not include any published state-of-the-art baselines because, to the best of our knowledge, all are either not publicly available or leverage proprietary data for training [24, 68, 69, 70].

Table 2: **Baseline results for the *de novo* molecule generation challenge.** The values in brackets indicate 99.9% confidence intervals upon bootstrapping (20,000 resamples).

	Top-1			Top-10		
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
Random chemical generation	0.00	**28.59 (28.33-28.84)**	0.07 (0.07 - 0.07)	0.00	25.72 (25.49-25.95)	0.10 (0.10 - 0.10)
SMILES Transformer	0.00	53.80 (52.95-54.61)	0.07 (0.07 - 0.08)	0.00	21.97 (21.79-22.16)	**0.17 (0.17 - 0.17)**
SELFIES Transformer	0.00	33.28 (33.00-33.57)	**0.10 (0.10 - 0.10)**	0.00	**21.84 (21.67-22.00)**	0.15 (0.15 - 0.15)
Bonus chemical formulae challenge						
SMILES Transformer	0.00	79.39 (78.64-80.08)	0.03 (0.03 - 0.04)	0.00	52.13 (51.45-52.81)	0.10 (0.09 - 0.10)
SELFIES Transformer	0.00	38.88 (38.57-39.20)	**0.08 (0.08 - 0.08)**	0.00	26.87 (26.66-27.11)	**0.13 (0.13 - 0.13)**
Random chemical generation	0.00	**21.11 (20.97-21.26)**	**0.08 (0.08 - 0.08)**	0.00	**18.25 (18.14-18.35)**	0.11 (0.11 - 0.11)

Table 3: **Baseline results for the molecule retrieval challenge.** The values in brackets indicate 99.9% confidence intervals upon bootstrapping (20,000 resamples).

	Hit rate @ 1 ↑	Hit rate @ 5 ↑	Hit rate @ 20 ↑	MCES @ 1 ↓
Random	0.37 (0.24-0.54)	2.01 (1.68-2.39)	8.22 (7.53-8.89)	30.81 (30.40-31.21)
DeepSets	1.47 (1.18-1.77)	6.21 (5.64-6.82)	19.23 (18.24-20.26)	25.11 (24.84-25.39)
Fingerprint FFN	2.54 (2.17-2.99)	7.59 (6.96-8.28)	20.00 (19.01-20.98)	24.66 (24.38-24.94)
DeepSets + Fourier features	5.24 (4.71-5.83)	12.58 (11.80-13.42)	28.21 (27.10-29.36)	22.13 (21.85-22.43)
MIST	**14.64 (13.82-15.54)**	**34.87 (33.69-36.10)**	**59.15 (57.89-60.39)**	**15.37 (15.12-15.62)**
Bonus chemical formulae challenge				
Random	3.06 (2.64-3.52)	11.35 (10.60-12.12)	27.74 (26.52-28.84)	13.87 (13.70-14.03)
DeepSets	4.42 (3.92-4.97)	14.46 (13.58-15.36)	30.76 (29.67-31.93)	15.04 (14.89-15.19)
Fingerprint FFN	5.09 (4.57-5.66)	14.69 (13.83-15.56)	31.97 (30.86-33.10)	14.94 (14.79-15.09)
DeepSets + Fourier features	6.56 (5.95-7.16)	16.46 (15.58-17.35)	33.46 (32.39-34.59)	14.14 (13.98-14.31)
MIST	**9.57 (8.88-10.30)**	**22.11 (21.10-23.13)**	**41.12 (39.98-42.34)**	**12.75 (12.59-12.91)**

Table 4: **Baseline results for the spectrum simulation challenge.** The values in brackets indicate 99.9% confidence intervals upon bootstrapping (20,000 resamples).

	Cosine Similarity ↑	Jensen-Shannon Similarity ↑	Hit Rate @ 1 ↑	Hit Rate @ 5 ↑	Hit Rate @ 20 ↑
Precursor m/z	0.15 (0.14-0.17)	0.59 (0.58-0.60)	0.38 (0.21-0.62)	1.72 (1.32-2.18)	7.17 (6.32-8.04)
FFN Fingerprint	0.25 (0.24-0.26)	0.69 (0.63-0.65)	8.44 (7.56-9.34)	21.43 (20.10-22.79)	38.57 (36.99-40.23)
GNN	0.19 (0.18-0.20)	0.64 (0.63-0.65)	3.95 (3.37-4.62)	11.92 (10.87-13.00)	26.27 (24.83-27.82)
FraGNNet	**0.52 (0.51-0.53)**	**0.91 (0.91-0.92)**	**46.64 (44.98-48.26)**	**72.56 (71.18-74.00)**	**83.58 (82.34-84.75)**
Bonus chemical formulae challenge					
Precursor m/z	-	-	2.09 (1.66-2.59)	8.52 (7.65-9.53)	22.65 (21.26-24.01)
FFN Fingerprint	-	-	7.62 (6.77-8.54)	22.70 (21.32-24.12)	44.12 (42.51-45.75)
GNN	-	-	3.63 (3.05-4.29)	13.55 (12.46-14.68)	33.77 (32.26-35.37)
FraGNNet	-	-	**31.93 (30.40-33.50)**	**63.20 (61.64-64.76)**	**82.70 (81.45-83.93)**

Molecule retrieval. The simplest baseline method for molecule retrieval, **Random**, sorts the candidate molecules $\mathcal{C}$ randomly. The second method, **Fingerprint FFN**, employs a feedforward neural network to predict the Morgan fingerprint of the target molecule. The candidates are then sorted based on their cosine similarity to the predicted fingerprint. Next, we evaluate **MIST**, a state-of-the-art deep learning approach, also based on fingerprint prediction. MIST assigns chemical subformulae to spectral peaks via energy-based modeling [22], then predicts a molecular fingerprint via a chemical formula-based transformer, and finally ranks the candidates by cosine similarity between the fingerprints [19]. Finally, we evaluate **DeepSets** [71]. The model processes spectra as sets of raw 2D peak representations, providing a complementary approach to FingerprintFFN and the state-of-the-art MIST which are based on alternative representations of spectra. **DeepSets + Fourier features** enhances DeepSets by using Fourier features enabling more accurate modeling of m/z values [20].

Spectrum simulation. We include three deep learning baseline models for the spectrum simulation task. The **molecular fingerprint (FFN Fingerprint)** model consists of a simple feedforward network on top of a fingerprint representation of the input molecule, inspired by [55]. The **graph neural network (GNN)** model, inspired by [56, 72], uses a variant of Graph Isomorphism Network augmented with edge features [73, 74] to process a 2D graph representation of the input molecule. Finally, state-of-the-art **FraGNNet** [23] uses combinatorial fragmentation and GNNs to parametrize

a probability distribution over fragments of the input molecule and their associated chemical formulae. The precise formula masses are used to map the distribution over formulae to a high resolution mass spectrum, without requiring binning. In addition, we include a trivial baseline **Precursor m/z** that simply predicts a single-peak spectrum by calculating the precursor m/z from the masses of the input molecule and the ionization adduct.

4.2 Baseline performance

We train and validate the performance of baseline methods on MassSpecGym. For the challenge of *de novo* generation (Table 2), we find that none of the baselines achieve an accuracy above zero, emphasizing the need for new method development. Additionally, our SMILES Transformer baseline does not outperform random generation of chemically valid graphs, highlighting the insufficiency of simplistic learning approaches in our generalization-demanding setup. For molecule retrieval (Table 3), the advanced MIST model significantly outperforms the simpler Fingerprint FFN and DeepSets baselines, suggesting strong gains from algorithmic development, which we posit as a driving force for MS/MS annotation with our benchmark. The same holds true for the spectrum simulation challenge (Table 4), where the advanced FraGNNet model demonstrates superior performance over simpler baselines. Nevertheless, the absolute metric scores still leave a substantial gap for future improvements.

5 Conclusions

In this work, we developed MassSpecGym, the first comprehensive and standardized benchmark for the discovery and identification of molecules from MS/MS spectra. MassSpecGym is based on our newly created largest open-source dataset of labeled tandem mass spectra and a standardization pipeline ensuring high data quality. We split the dataset using our novel generalization-demanding splitting technique, enabling robust evaluation of molecular identification and discovery. We evaluated a series of baseline methods and demonstrated that the annotation of MS/MS spectra remains a highly unsolved problem. To address this, we provide MassSpecGym as a public resource with a user-friendly interface requiring minimal domain expertise for the submission and evaluation of new machine learning models.

Our future work has two main directions. First, we plan to continuously update MassSpecGym with new public and in-house MS/MS data, potentially incorporating simulated spectra or additional data modalities, such as EI spectra. We also aim to expand the scope of challenges to include tasks such as molecular networking, a prominent technique in the field that focuses on clustering spectra of structurally related molecules rather than predicting individual molecules. Second, by progressively enhancing the MassSpecGym ecosystem with more advanced methods, we intend to transform it into a hub for state-of-the-art models in MS/MS spectra annotation. This will empower machine learning researchers to make rapid progress in developing innovative models, a particularly crucial focus given the historically limited collaboration between mass spectrometry experts and AI specialists. As a consequence, many well-established machine learning paradigms, such as generating molecular graphs via diffusion models or applying domain adaptation techniques across different mass spectrometry systems, remain largely unexplored. Furthermore, by providing a user-friendly interface to run these models, we aim to make cutting-edge algorithms readily accessible to life scientists interested in annotating their mass spectra. We believe that MassSpecGym will play a pivotal role in fostering the development of next-generation machine learning methods, ultimately driving significant progress across the biomedical and chemical sciences.

Acknowledgments and Disclosure of Funding

The idea of this benchmark set was conceived at the Dagstuhl Seminar #24181 “Computational Metabolomics: Towards Molecules, Models, and their Meaning”. We are grateful for the funding and support provided by the Leibniz Center for Informatics.

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140) and by the Technology Agency of the Czech Republic through the project RETEMED (TN02000122). This work was also co-funded by the European Union (ERC FRONTIER No. 101097822; ELIAS No. 101120237). 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. SH and AK are funded through Award R35GM148219 by the NIGMS of the National Institutes of Health. TP was 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. AY was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) scholarship and a Vector Institute research grant. LZ was supported by NSERC. FW was supported by Alberta Machine Intelligence Institute (AMII), and NSERC grant. RG was was supported by NSERC, AMII. DSW was supported by NSERC, and Genome Canada, Genome British Columbia and Genome Alberta. HR was supported by NSERC and the Canada Research Chair (CRC) Program. FK, KD and SB are supported by Deutsche Forschungsgemeinschaft (BO 1910/23). FK and SB are supported by the Ministry for Economics, Sciences and Digital Society of Thuringia (Framework ProDigital, DigLeben-5575/10-9). NAH and SB are supported by the Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft (TMWWDG) with funds from the European Union as part of the European Regional Development Fund (ERDF, 2023 VFE 0003). JH and WB acknowledge support by the University of Antwerp Research Fund. FH was supported by Deutsche Forschungsgemeinschaft (DFG, 528775510). LL was supported by NSF Award 2239869. CB was supported by the Czech Academy of Sciences PPLZ fellowship number L200552251.

Competing interests

RS and TP are co-founders of the company mzio GmbH, which develops technologies related to mass spectrometry data processing. SB, KD and ML are co-founders of Bright Giant GmbH. JJJvdH is member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy and consults for Corteva Agriscience, Indianapolis, IN, USA.

References

[1] Veronica Termopoli, Elena Torrisi, Giorgio Famiglini, Pierangela Palma, Giovanni Zappia, Achille Cappiello, Gregory W. Vandergrift, Misha Zvekic, Erik T. Krogh, and Chris G. Gill. Mass spectrometry based approach for organic synthesis monitoring. *Analytical Chemistry*, 91 (18):11916–11922, 2019. doi: 10.1021/acs.analchem.9b02681. URL `https://doi.org/10.1021/acs.analchem.9b02681`. PMID: 31403767.

[2] Umakant Sahu, Elodie Villa, Colleen R. Reczek, Zibo Zhao, Brendan P. O'Hara, Michael D. Torno, Rohan Mishra, William D. Shannon, John M. Asara, Peng Gao, Ali Shilatifard, Navdeep S. Chandel, and Issam Ben-Sahra. Pyrimidines maintain mitochondrial pyruvate oxidation to support de novo lipogenesis. *Science*, 383(6690):1484–1492, 2024. doi: 10.1126/science.adh2771. URL `https://www.science.org/doi/abs/10.1126/science.adh2771`.

[3] Juan Carlos Alarcon-Barrera, Sarantos Kostidis, Alejandro Ondo-Mendez, and Martin Giera. Recent advances in metabolomics analysis for early drug development. *Drug discovery today*, 27(6):1763–1773, 2022. doi: 10.1016/j.drudis.2022.02.018.

[4] Mamas Mamas, Warwick B Dunn, Ludwig Neyses, and Royston Goodacre. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of toxicology*, 85:5–17, 2011. doi: 10.1007/s00204-010-0609-6.

[5] Juliane Hollender, Emma L. Schymanski, Heinz P. Singer, and P. Lee Ferguson. Nontarget screening with high resolution mass spectrometry in the environment: Ready to go? *Environmental Science & Technology*, 51(20):11505–11512, Oct 2017. ISSN 0013-936X. doi: 10.1021/acs.est.7b02184. URL `https://doi.org/10.1021/acs.est.7b02184`.

[6] Jillian Romsdahl, Adriana Blachowicz, Yi-Ming Chiang, Kasthuri Venkateswaran, and Clay CC Wang. Metabolomic analysis of aspergillus niger isolated from the international space station reveals enhanced production levels of the antioxidant pyranonigrin a. *Frontiers in Microbiology*, 11:529292, 2020. doi: 10.3389/fmicb.2020.00931.

[7] Saleh Alseekh, Asaph Aharoni, Yariv Brotman, Kévin Contrepois, John D'Auria, Jan Ewald, Jennifer C. Ewald, Paul D. Fraser, Patrick Giavalisco, Robert D. Hall, Matthias Heinemann, Hannes Link, Jie Luo, Steffen Neumann, Jens Nielsen, Leonardo Perez de Souza, Kazuki Saito, Uwe Sauer, Frank C. Schroeder, Stefan Schuster, Gary Siuzdak, Aleksandra Skirycz, Lloyd W. Sumner, Michael P. Snyder, Huiru Tang, Takayuki Tohge, Yulan Wang, Weiwei Wen, Si Wu, Guowang Xu, Nicola Zamboni, and Alisdair R. Fernie. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature Methods*, 18(7):747–756, Jul 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01197-1. URL https://doi.org/10.1038/s41592-021-01197-1.

[8] Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S. Mehta, Gert Wohlgemuth, Dinesh Kumar Barupal, Megan R. Showalter, Masanori Arita, and Oliver Fiehn. Identification of small molecules using accurate mass ms/ms search. *Mass Spectrometry Reviews*, 37(4):513–532, 2018. doi: https://doi.org/10.1002/mas.21535. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mas.21535.

[9] Juan Carlos Alarcon-Barrera, Sarantos Kostidis, Alejandro Ondo-Mendez, and Martin Giera. Recent advances in metabolomics analysis for early drug development. *Drug Discovery Today*, 27(6):1763–1773, 2022. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2022.02.018. URL https://www.sciencedirect.com/science/article/pii/S1359644622000769.

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

[11] Beate I. Escher, Heather M. Stapleton, and Emma L. Schymanski. Tracking complex mixtures of chemicals in our changing environment. *Science*, 367(6476):388–392, 2020. doi: 10.1126/science.aay6636. URL https://www.science.org/doi/abs/10.1126/science.aay6636.

[12] Renaud Prudent, D. Allen Annis, Peter J. Dandliker, Jean-Yves Ortholand, and Didier Roche. Exploring new targets and chemical space with affinity selection-mass spectrometry. *Nature Reviews Chemistry*, 5(1):62–71, Jan 2021. ISSN 2397-3358. doi: 10.1038/s41570-020-00229-2. URL https://doi.org/10.1038/s41570-020-00229-2.

[13] Shi Qiu, Ying Cai, Hong Yao, Chunsheng Lin, Yiqiang Xie, Songqi Tang, and Aihua Zhang. Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduction and Targeted Therapy*, 8(1):132, Mar 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01399-3. URL https://doi.org/10.1038/s41392-023-01399-3.

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

[15] Niek F. de Jonge, Kevin Mildau, David Meijer, Joris J. R. Louwen, Christoph Bueschl, Florian Huber, and Justin J. J. van der Hooft. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics*, 18 (12):103, Dec 2022. ISSN 1573-3890. doi: 10.1007/s11306-022-01963-y. URL https://doi.org/10.1007/s11306-022-01963-y.

[16] Parisa Bayat, Denis Lesage, and Richard B. Cole. Tutorial: Ion activation in tandem mass spectrometry using ultra-high resolution instrumentation. *Mass Spectrometry Reviews*, 39(5-6):680–702, 2020. doi: https://doi.org/10.1002/mas.21623. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mas.21623.

[17] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16 (4):299–302, Apr 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0344-8. URL https://doi.org/10.1038/s41592-019-0344-8.

[18] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. Cfm-id 4.0: More accurate esi-ms/ms spectral prediction and compound identification. *Analytical Chemistry*, 93(34):11692–11700, Aug 2021. ISSN 0003-2700. doi: 10.1021/acs.analchem.1c01465. URL https://doi.org/10.1021/acs.analchem.1c01465.

[19] Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J Xavier, and Connor W Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, 2023. doi: https://doi.org/10.1038/s42256-023-00708-3.

[20] Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Emergence of molecular structures from repository-scale self-supervised learning on tandem mass spectra. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2023-kss3r-v2.

[21] Samuel Goldman, Janet Li, and Connor W. Coley. Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks, January 2024. URL http://arxiv.org/abs/2304.13136.

[22] Samuel Goldman, Jiayi Xin, Joules Provenzano, and Connor W. Coley. Mist-cf: Chemical formula inference from tandem mass spectra. *Journal of Chemical Information and Modeling*, 64(7):2421–2431, Apr 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01082. URL https://doi.org/10.1021/acs.jcim.3c01082.

[23] Adamo Young, Fei Wang, David Wishart, Bo Wang, Hannes Röst, and Russ Greiner. FraGNNet: A Deep Probabilistic Model for Mass Spectrum Prediction, April 2024. URL http://arxiv.org/abs/2404.02360.

[24] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. Msnovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022. doi: 10.1038/s41592-022-01486-3.

[25] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 25549–25562. PMLR, 2023. URL https://proceedings.mlr.press/v202/murphy23a.html.

[26] Richard Overstreet, Ethan King, Grady Clopton, Julia Nguyen, and Danielle Ciesielski. Qcgn2oms2: a graph neural net for high resolution mass spectra prediction. *Journal of Chemical Information and Modeling*, 64(15):5806–5816, Aug 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.4c00446. URL https://doi.org/10.1021/acs.jcim.4c00446.

[27] Richard Licheng Zhu and Eric Jonas. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical Chemistry*, 95(5):2653–2663, Feb 2023. ISSN 0003-2700. doi: 10.1021/acs.analchem.2c02093. URL https://doi.org/10.1021/acs.analchem.2c02093.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1264. URL https://doi.org/10.18653/v1/d16-1264.

[30] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016. URL http://arxiv.org/abs/1606.01540.

[31] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N. Gomez, Debora S. Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16990–17017. PMLR, 2022. URL https://proceedings.mlr.press/v162/notin22a.html.

[32] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan J. Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html.

[33] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *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/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

[34] Corinna Brungs, Robin Schmid, Steffen Heuckeroth, Aninda Mazumdar, Matúš Drexler, Pavel Šácha, Pieter C Dorrestein, Daniel Petras, Louis-Felix Nothias, Radim Nencka, et al. Efficient generation of open multi-stage fragmentation mass spectral libraries. 2024. doi: 10.26434/chemrxiv-2024-l1tqh.

[35] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology*, 39(4):462–471, 2021. doi: 10.1038/s41587-020-0740-8.

[36] Mingxun Wang, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, and et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8):828–837, Aug 2016. ISSN 1546-1696. doi: 10.1038/nbt.3597. URL https://doi.org/10.1038/nbt.3597.

[37] MoNA. Massbank of north america. https://mona.fiehnlab.ucdavis.edu/. URL https://mona.fiehnlab.ucdavis.edu/.

[38] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714, 2010. doi: 10.1002/jms.1777.

[39] Stephen Stein. Mass spectral reference libraries: an ever-expanding resource for chemical identification, 2012.

[40] Wout Bittremieux, Mingxun Wang, and Pieter C Dorrestein. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics*, 18(12):94, 2022. doi: https://doi.org/10.1007/s11306-022-01947-y.

[41] mzCloud. mzcloud. https://www.mzcloud.org/. URL https://www.mzcloud.org/.

[42] Hartmuth C Kolb, MG Finn, and K Barry Sharpless. Click chemistry: diverse chemical function from a few good reactions. *Angewandte Chemie International Edition*, 40(11):2004–2021, 2001. doi: https://doi.org/10.1002/1521-3773(20010601)40:11<2004::AID-ANIE2004>3.0.CO;2-5. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1521-3773%2820010601%2940%3A11%3C2004%3A%3AAID-ANIE2004%3E3.0.CO%3B2-5.

[43] Fleming Kretschmer, Jan Seipp, Marcus Ludwig, Gunnar W Klau, and Sebastian Boecker. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv*, pages 2023–03, 2023. doi: doi.org/10.1101/2023.03.27.534311.

[44] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, Russell Greiner, and David S. Wishart. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, 8(1):61, Nov 2016. ISSN 1758-2946. doi: 10.1186/s13321-016-0174-y. URL https://doi.org/10.1186/s13321-016-0174-y.

[45] Critical assessment of small molecule identification. casmi. http://www.casmi-contest.org/2022/index.shtml, 2022.

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

[47] Daniel P. Demarque, Antonio E. M. Crotti, Ricardo Vessecchi, João L. C. Lopes, and Norberto P. Lopes. Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. *Nat. Prod. Rep.*, 33:432–455, 2016. doi: 10.1039/C5NP00073D. URL http://dx.doi.org/10.1039/C5NP00073D.

[48] Maria Vinaixa, Emma L. Schymanski, Steffen Neumann, Miriam Navarro, Reza M. Salek, and Oscar Yanes. 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. ISSN 0165-9936. doi: https://doi.org/10.1016/j.trac.2015.09.005. URL https://www.sciencedirect.com/science/article/pii/S0165993615300832.

[49] Emma L. Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, 48(4):2097–2098, Feb 2014. ISSN 0013-936X. doi: 10.1021/es5002105. URL https://doi.org/10.1021/es5002105.

[50] Sebastian Böcker and Kai Dührkop. Fragmentation trees reloaded. *Journal of cheminformatics*, 8:1–26, 2016. doi: 10.1186/s13321-016-0116-8.

[51] Shipei Xing, Sam Shen, Banghua Xu, Xiaoxiao Li, and Tao Huan. Buddy: molecular formula discovery via bottom-up ms/ms interrogation. *Nature Methods*, 20(6):881–890, Jun 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01850-x. URL https://doi.org/10.1038/s41592-023-01850-x.

[52] Tomáš Pluskal, Taisuke Uehara, and Mitsuhiro Yanagida. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, ms/ms fragmentation, heuristic rules, and isotope pattern matching. *Analytical Chemistry*, 84(10):4396–4403, May 2012. ISSN 0003-2700. doi: 10.1021/ac3000418. URL https://doi.org/10.1021/ac3000418.

[53] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[54] PubChem. PubChem. https://pubchem.ncbi.nlm.nih.gov/docs/statistics/. URL https://pubchem.ncbi.nlm.nih.gov/docs/statistics.

[55] Jennifer N Wei, David Belanger, Ryan P Adams, and D Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science*, 5(4):700–708, 2019. doi: 10.1021/acscentsci.9b00085.

[56] Hao Zhu, Liping Liu, and Soha Hassoun. Using graph neural networks for mass spectrometry prediction. *arXiv preprint arXiv:2010.04661*, 2020. doi: 10.48550/arXiv.2010.04661.

[57] Adamo Young, Hannes Röst, and Bo Wang. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416, April 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00816-8. URL https://www.nature.com/articles/s42256-024-00816-8.

[58] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12(1):12, Feb 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-0416-x. URL https://doi.org/10.1186/s13321-020-0416-x.

[59] Florian Huber, Stefan Verhoeven, Christiaan Meijer, Hanno Spreeuw, Efraín Manuel Villanueva Castilla, Cunliang Geng, Justin J. j. van der Hooft, Simon Rogers, Adam Belloum, Faruk Diblen, and Jurriaan H. Spaaks. matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, 5(52):2411, 2020. doi: 10.21105/joss.02411. URL https://doi.org/10.21105/joss.02411.

[60] Niek F de Jonge, Helge Hecht, Justin JJ van der Hooft, and Florian Huber. Reproducible ms/ms library cleaning pipeline in matchms. *ChemRxiv preprint*, 2023. doi: doi:10.26434/chemrxiv-2023-l44cm.

[61] Kai Dührkop, Marcus Ludwig, Marvin Meusel, and Sebastian Böcker. Faster mass decomposition. In Aaron E. Darling and Jens Stoye, editors, *Algorithms in Bioinformatics - 13th International Workshop, WABI 2013, Sophia Antipolis, France, September 2-4, 2013. Proceedings*, volume 8126 of *Lecture Notes in Computer Science*, pages 45–58. Springer, 2013. doi: 10.1007/978-3-642-40453-5_5. URL https://doi.org/10.1007/978-3-642-40453-5_5.

[62] Volker D. Hähnke, Sunghwan Kim, and Evan E. Bolton. Pubchem chemical structure standardization. *Journal of Cheminformatics*, 10(1):36, Aug 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0293-8. URL https://doi.org/10.1186/s13321-018-0293-8.

[63] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Res*, 51(D1):D1373–D1380, January 2023.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762.

[65] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1162. URL https://doi.org/10.18653/v1/p16-1162.

[66] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988. URL https://api.semanticscholar.org/CorpusID:5445756.

[67] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.*, 1(4):45024, 2020. doi: 10.1088/2632-2153/ABA947. URL https://doi.org/10.1088/2632-2153/aba947.

[68] T. Butler, A. Frandsen, R. Lightheart, B. Bargh, T. Kerby, K. West, and et al. Ms2mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv*, 2023. doi: 10.26434/chemrxiv-2023-vsmpx-v4.

[69] Litsa E, Chenthamarakshan V, Das P, and Kavraki L. Spec2mol: An end-to-end deep learning framework for translating ms/ms spectra to de-novo molecules. *ChemRxiv*, 2021. doi: doi: 10.26434/chemrxiv-2021-6rdh6.

[70] David Elser, Florian Huber, and Emmanuel Gaquerel. Mass2smiles: deep learning based fast prediction of structures and functional groups directly from high-resolution ms/ms spectra. *bioRxiv*, 2023. doi: 10.1101/2023.07.06.547963. URL `https://www.biorxiv.org/content/early/2023/07/08/2023.07.06.547963`.

[71] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3391–3401, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html`.

[72] Xinmeng Li, Yan Zhou Chen, Apurva Kalia, Hao Zhu, Li-ping Liu, and Soha Hassoun. An ensemble spectral prediction (esp) model for metabolite annotation. *Bioinformatics*, 40(8): btae490, 2024. doi: https://doi.org/10.1093/bioinformatics/btae490.

[73] Keyulu Xu*, Weihua Hu*, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? September 2018. URL `https://openreview.net/forum?id=ryGs6iA5Km`.

[74] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. doi: 10.48550/arXiv.1903.02428.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Conclusions.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not expect any negative social impact from our benchmark for the analytical chemistry problem.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]
(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code, data and instructions are available at `https://github.com/pluskal-lab/MassSpecGym`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All the training details related to the benchmark (e.g., data splits) are discussed in the text and publicly available as part of the data. The training details related to the models are discussed in the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We did not run training experiments multiple times due to computational demands. However, we keep random seeds fixed for reproducibility.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The information is provided in the supplemental material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] We are not using any commercially-licensed assets.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All the new assets are publicly available through `https://github.com/pluskal-lab/MassSpecGym`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our mass spectrometry data does not contain personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

MassSpecGym: Supplementary information

Roman Bushuiev[1,2], **Anton Bushuiev**[2], **Niek F. de Jonge**[3], **Adamo Young**[4], **Fleming Kretschmer**[5], **Raman Samusevich**[1,2], **Janne Heirman**[6], **Fei Wang**[7,8], **Luke Zhang**[9], **Kai Dührkop**[5], **Marcus Ludwig**[10], **Nils A. Haupt**[5], **Apurva Kalia**[11], **Corinna Brungs**[1], **Robin Schmid**[1], **Russell Greiner**[7,8], **Bo Wang**[4], **David S. Wishart**[7,12], **Li-Ping Liu**[11], **Juho Rousu**[13], **Wout Bittremieux**[6], **Hannes Rost**[9], **Tytus D. Mak**[14], **Soha Hassoun**[11,15], **Florian Huber**[16], **Justin J.J. van der Hooft**[3,17], **Michael A. Stravs**[18], **Sebastian Böcker**[5], **Josef Sivic**[2], **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, [3]Bioinformatics Group, Wageningen University & Research, [4]Department of Computer Science, University of Toronto, [5]Chair for Bioinformatics, Institute for Computer Science, Friedrich Schiller University Jena, [6]Department of Computer Science, University of Antwerp, [7]Department of computing science, University of Alberta, [8]Alberta Machine Intelligence Institute, [9]Department of Molecular Genetics, University of Toronto, [10]Bright Giant GmbH, [11]Department of Computer Science, Tufts University, [12]Department of Biological Sciences, University of Alberta, [13]Department of Computer Science, Aalto University, [14]Mass Spectrometry Data Center, National Institute of Standards and Technology, [15]Department of Chemical and Biological Engineering, Tufts University, [16]Centre for Digitalisation and Digitality, University of Applied Sciences Düsseldorf, [17]Department of Biochemistry, University of Johannesburg, [18]Eawag: Swiss Federal Institute of Aquatic Science and Technology

Contents

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

1 Dataset and code access

1.1 Availability

Following the NeurIPS Dataset and Benchmark Track guidelines, we have made our dataset publicly available under the MIT license. The dataset and its Croissant metadata record can be accessed through the Hugging Face Hub. Furthermore, we have made the code for dataset construction, analysis, reproduction of all our experiments, and evaluation of new models available on GitHub under the MIT license. Figure 1 provides an overview of the MassSpecGym infrastructure. We bear all responsibility in case of any violation of rights.

- MassSpecGym dataset:
  `https://huggingface.co/datasets/roman-bushuiev/MassSpecGym`.
- MassSpecGym code:
  `https://github.com/pluskal-lab/MassSpecGym`.

1.2 Variable list

Table 1 outlines the structure of our MassSpecGym dataset and provides the list of all variables along with their descriptions. Figure 2 explains the key relationship between the main variables of individual samples: multiple spectra with different metadata, measured under varying experimental setups, may be annotated with the same molecule.

1.3 Dataset Applications

The primary goal of MassSpecGym is to identify the most effective machine learning models for the annotation of MS/MS spectra with molecular structures. Our benchmark, which defines a generalization-demanding data split and practically-motivated evaluation metrics, ensures that models performing well on MassSpecGym can be effectively applied to real-world, unannotated data. Ultimately, MassSpecGym paves the way for new biological and chemical discoveries by stimulating advances in the annotation of MS/MS spectra. On the other hand, *de novo* molecule generation serves

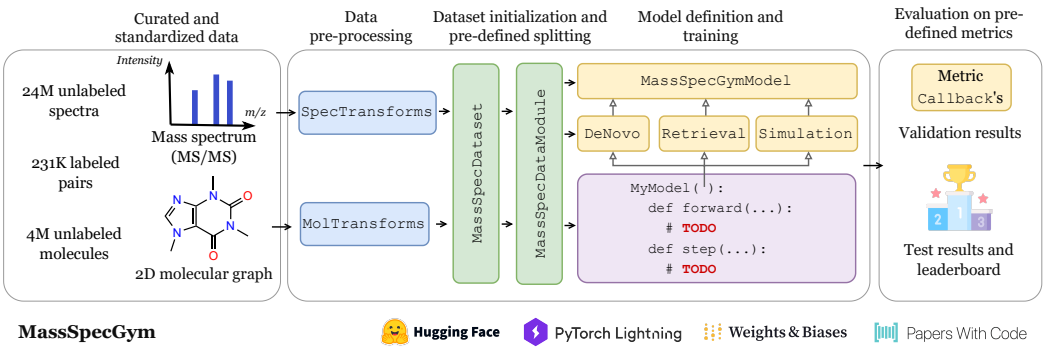

Figure 1: **MassSpecGym enables a standardized and user-friendly evaluation of machine learning methods for MS/MS annotation via an easily extendable modular interface.** The dataset can be loaded, preprocessed, split, and utilized for training, evaluation, and metric logging through the prepared codebase. To develop and evaluate a new model, a user only needs to implement a forward pass with custom prediction logic. Colored blocks represent classes in our codebase (`https://github.com/pluskal-lab/MassSpecGym`). Arrows with empty heads represent subclass inheritance, while arrows with bold heads conceptually show the flow from the dataset to the evaluation metrics.

Table 1: **Overview of all variables present in the MassSpecGym dataset.** n denotes the number of signals in a spectrum. Floating point variables were rounded to four decimal places for computing the number of unique entries. The key variables in the MassSpecGym dataset are `mzs` and `intensities` representing an input spectrum, and `smiles`, representing the target molecule.

Variable	Description	Data type	Num. unique values	Example
`identifier`	Unique entry identifier	string	231,104	MassSpecGymID0088683
`mzs`	Array of spectrum m/z values	$n \times$ float	231,104	[55.0542, 57.0699, . . . , 238.0995]
`intensities`	Array of spectrum intensities	$n \times$ float	231,104	[0.0240, 1.0, . . . , 0.5356]
`smiles`	SMILES string of molecule	string	31,602	CCCCOCN(C1=C(C=C. . . CCl
`inchikey`	2D InChI key	string	28,929	HKPHPIREJKHECO
`formula`	Chemical formula of molecule	string	17,634	C17H26ClNO2
`precursor_formula`	Chemical formula of precursor ion	string	21,653	C17H27ClNO2
`parent mass`	Mass of molecule	float	32,228	311.1652
`precursor_mz`	M/z of precursor ion	float	32,275	312.1725
`adduct`	Ionization adduct	string	2	[M+H]+
`instrument_type`	Type of MS instrument	string	2	Orbitrap
`collision_energy`	Energy of CID fragmentation	float	9,737	30.0
`fold`	Split fold which entry belongs to	string	3	train
`simulation_challenge`	Entry is used for simulation challenge	boolean	2	True

as a benchmark for evaluating novel generative modeling algorithms. With the rapid advancements in the field of deep generative modeling, our benchmark provides a rigorous case study to assess their capabilities. Similarly, the molecule retrieval challenge acts as a benchmark for evaluating information retrieval algorithms. The spectrum simulation benchmark has the potential to provide insights into the mechanisms of MS/MS fragmentation, thereby deepening our understanding of the underlying processes in analytical chemistry.

1.4 Target Audiences

MassSpecGym aims to democratize and popularize the challenge of annotating molecular structures from MS/MS spectra. Our platform targets two primary audiences:

- Machine learning community: Researchers and developers specializing in creating new machine learning algorithms, who may not necessarily have a background in mass spectrometry or other fields of chemistry.
- Metabolomics community: Scientists and professionals specializing in mass spectrometry-related fields, such as natural product chemists, who can utilize models trained on MassSpecGym and apply insights from the MassSpecGym results to their research problems.

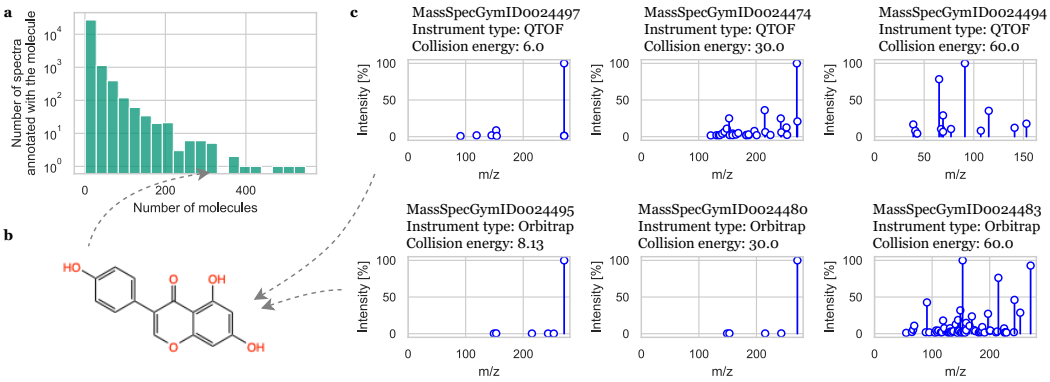

Figure 2: **Same measured molecule may result in different MS/MS spectra under different mass spectrometry measurement conditions. a,** The distribution of the number of spectra corresponding to the same molecule in the MassSpecGym dataset. **b,** Example of a molecule annotating 321 spectra in the dataset. **c,** Examples of six spectra annotated with the molecule shown in figure **b**: different instrument types and collision energies lead to different spectra. Higher collision energies typically lead to richer fragmentation of a molecule, resulting in a higher number of peaks in the spectrum.

MassSpecGym streamlines the cleaning and pre-processing of data, removing the necessity for specialized mass spectrometry expertise. This makes the platform accessible to a wider audience, including those who may not have access to extensive computational resources for collecting and cleaning large-scale data (e.g., large unlabeled datasets provided in our benchmark).

2 MassSpecGym benchmark construction

This section provides details on the construction of the MassSpecGym benchmark that are not covered in the main text.

2.1 Definition of challenges

Morgan fingerprint. A Morgan fingerprint [1] is a set of structural features of a molecule. For machine learning purposes, it is typically represented as a bit vector, where each bit indicates the presence of a specific fragment in the molecule. The dimensionality of a Morgan fingerprint refers to the number of bits in the vector, usually set to a fixed size such as 2048 or 4096 bits. An additional parameter of the Morgan fingerprint is its radius. The radius specifies the number of steps (or bonds) to consider from each atom when generating the fingerprint, with a common choice being a radius of 2, which considers all the neighbors up to two bonds away for each atom. This helps capture the local structural environment around each atom within the molecule. Further, we denote a Morgan fingerprint of a molecular graph G, representing a sets of it structural features, as $fingerprint(G)$.

Tanimoto Similarity Tanimoto similarity measures the similarity between two molecular graphs G_1, G_2 using their Morgan fingerprints. It is defined as the intersection over union of the two corresponding fingerprints:

$$Tanimoto(G_1, G_2) = \frac{|fingerprint(G_1) \cap fingerprint(G_2)|}{|fingerprint(G_1) \cup fingerprint(G_2)|}. \quad (1)$$

The Tanimoto similarity score ranges from 0 to 1, with 1 indicating identical structures. However, in rare cases, non-identical structures can also produce a similarity score of 1 due to fingerprint collisions. A similarity value above 0.85 has been previously used in several studies as a threshold for identifying compounds with similar activity [2].

MCES distance. To measure similarity between two molecular structures, additional to the Tanimoto similarity, the Maximum Common Edge Subgraph (MCES) distance is used on the molecular graphs [3]. MCES distance can be understood as an edit distance on graphs, representing the minimum

number of edges that have to be removed for both graphs to be isomorphic, ignoring singleton nodes. This measure of molecular similarity provides better interpretability than the Tanimoto similarity and more accurately reflects the biochemical modifications of molecules. Formally, given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where V_1, V_2 are the sets of vertices and E_1, E_2 are the sets of edges, the maximum common edge subgraph $G_c = (V_c, E_c)$, $V_c \subseteq V_i$, $E_c \subseteq E_i$ for both $i \in \{1, 2\}$ is a graph that maximizes $|E_c|$. The MCES distance is then defined as

$$MCES(G_1, G_2) = |E_1| + |E_2| - 2|E_c|, \tag{2}$$

which is minimized by the maximum common edge subgraph. We use the myopicMCES implementation (`https://github.com/AlBi-HHU/myopic-mces`), which computes the exact distance below a specified threshold and a guaranteed lower bound above; the distance is weighted by bond order. Unless otherwise stated, we use a threshold of 15 and enable the `always_stronger_bound` option (default) which always computes a stronger lower bound first despite higher computational demand.

2.2 Data collection

Public repositories. The mass spectral data is composed of various LC-MS/MS databases and datasets, which were downloaded on May 27th, 2024, from:

- MoNA [4] downloaded from https://mona.fiehnlab.ucdavis.edu/downloads
- Massbank [5] downloaded from https://github.com/MassBank/MassBank-data/releases
- GNPS [6] The following libraries were downloaded from https://gnps-external.ucsd.edu/gnpslibrary
  - BERKELEY-LAB.mgf
  - BILELIB19.mgf
  - BIRMINGHAM-UHPLC-MS-NEG.mgf
  - BIRMINGHAM-UHPLC-MS-POS.mgf
  - BMDMS-NP.mgf
  - CASMI.mgf
  - DRUGS-OF-ABUSE-LIBRARY.mgf
  - ECG-ACYL-AMIDES-C4-C24-LIBRARY.mgf
  - ECG-ACYL-ESTERS-C4-C24-LIBRARY.mgf
  - GNPS-COLLECTIONS-MISC.mgf
  - GNPS-COLLECTIONS-PESTICIDES-NEGATIVE.mgf
  - GNPS-COLLECTIONS-PESTICIDES-POSITIVE.mgf
  - GNPS-D2-AMINO-LIPID-LIBRARY.mgf
  - GNPS-EMBL-MCF.mgf
  - GNPS-FAULKNERLEGACY.mgf
  - GNPS-IOBA-NHC.mgf
  - GNPS-LIBRARY.mgf
  - GNPS-MSMLS.mgf
  - GNPS-NIH-CLINICALCOLLECTION1.mgf
  - GNPS-NIH-CLINICALCOLLECTION2.mgf
  - GNPS-NIH-NATURALPRODUCTSLIBRARY.mgf
  - GNPS-NIH-NATURALPRODUCTSLIBRARY_ROUND2_NEGATIVE.mgf
  - GNPS-NIH-NATURALPRODUCTSLIBRARY_ROUND2_POSITIVE.mgf
  - GNPS-NIH-SMALLMOLECULEPHARMACOLOGICALLYACTIVE.mgf
  - GNPS-NIST14-MATCHES.mgf
  - GNPS-NUTRI-METAB-FEM-NEG.mgf
  - GNPS-NUTRI-METAB-FEM-POS.mgf
  - GNPS-PRESTWICKPHYTOCHEM.mgf
  - GNPS-SAM-SIK-KANG-LEGACY-LIBRARY.mgf
  - GNPS-SCIEX-LIBRARY.mgf

- GNPS-SELLECKCHEM-FDA-PART1.mgf
- GNPS-SELLECKCHEM-FDA-PART2.mgf
- HCE-CELL-LYSATE-LIPIDS.mgf
- HMDB.mgf
- IQAMDB.mgf
- LDB_NEGATIVE.mgf
- LDB_POSITIVE.mgf
- MIADB.mgf
- MMV_NEGATIVE.mgf
- MMV_POSITIVE.mgf
- PNNL-LIPIDS-NEGATIVE.mgf
- PNNL-LIPIDS-POSITIVE.mgf
- PSU-MSMLS.mgf
- RESPECT.mgf
- SUMNER.mgf
- UM-NPDC.mgf

In-house data. In-house spectral libraries were acquired for four different compound libraries, including the Bioactive Compound Library and the 5k Scaffold Library from MedChemExpress, the NIH NPAC ACONN collection of Natural Products from NIH/NCATS, and the Alpha-helix Peptidomimetic Library from OTAVAchemicals, resulting in almost 20,000 unique compounds. Up to 10 compounds were pooled in one well of 96 or 384 well plates and diluted to reach a concentration between 5-20 µM. A flow injection method was performed by a Vanquish Duo UHPLC system coupled to an Orbitrap ID-X instrument. 2 µL injection volume was used. The m/z range for MS^1 was set to 115 to 2000 with a resolution of 30,000. The three most intense ions were picked by data-dependent acquisition for MS^2 experiments with a resolution of 15,000. Detailed information about the data acquisition can be found in [7]. We used the following files downloaded from `https://zenodo.org/records/11163381`.

- 20231031_nihnp_library_neg_all_lib_MSn.mgf
- 20231130_mcescaf_library_neg_all_lib_MSn.mgf
- 20231130_otavapep_library_neg_all_lib_MSn.mgf
- 20240411_mcebio_library_neg_all_lib_MSn.mgf
- 20231031_nihnp_library_pos_all_lib_MSn.mgf
- 20231130_mcescaf_library_pos_all_lib_MSn.mgf
- 20231130_otavapep_library_pos_all_lib_MSn.mgf
- 20240411_mcebio_library_pos_all_lib_MSn.mgf

2.3 Data cleaning

To clean the collected dataset, we applied spectrum-based filtering and metadata-based filtering using the matchms package [8], as well as spectral quality-based filtering utilizing SIRIUS software [9] and the NIST20 spectral library [10]. The following subsections describe the individual steps. The reduction in the number of mass spectra after each filtering step is summarized in Table 2, following the matchms cleaning report format.

Spectrum-based filtering. Noise was detected and removed by first searching for multiple identical intensity values and then using the associated intensity value multiplied by 2 as a minimum intensity threshold. Spectra that contained more than 300 or less than 1 fragment after noise removal were discarded. Spectra with patterns that suggested that they were stored as profile spectra were also removed. It is not uncommon to depose data in multiple databases; therefore, duplicates, i.e. spectra with identical annotation and a cosine score of 1.0, were removed. We removed all entries with precursor m/z values above 1,000 Da. Additionally, we dropped all spectra containing at least one signal with an m/z value above the precursor m/z value + 3 and an intensity value above 20% of the highest intensity within the spectrum.

Metadata-based filtering. The spectra from public libraries have a high diversity of metadata formats and often have missing or incorrect metadata. The metadata was largely processed using default settings as described in [11] with the full workflow being provided on `https://github.com/pluskal-lab/MassSpecGym/blob/main/notebooks/dataset_construction`.

First all merged spectra in the MS/MS library from Brungs et al. [7] were removed.

Metadata fields were harmonized to have the same field names. Metadata in the wrong field was corrected. Annotations were completed, to have a SMILES, InChI and InChIKey. Missing parent masses were derived from the SMILES mass and missing adducts were derived from the precursor m/z and parent mass. Only spectra where the parent mass calculated from the adduct and precursor m/z matched the given parent mass and SMILES mass were stored.

Annotations were derived from the compound name by searching on PubChem. Spectra were removed if the mono-isomeric mass of the SMILES did not match the parent mass given in the metadata. Annotations which corresponded to a charged molecule were removed. The formula and precursor formula were derived from the SMILES. Additionally, entries with molecules containing rare chemical elements `Sn` and `Al` were removed.

For spectra to be technically comparable we limited the data to LC-MS/MS spectra measured in positive electrospray ionization (ESI) mode, attributable to Orbitrap-type (Orbitrap, ITFT, QFT) or QTOF-type instruments (see below). We further only retained spectra of two adducts, [M+H]+ or [M+Na]+, to reduce additional biases.

Next, we removed all structures that are disconnected (i.e. containing a '.' within the SMILES string). These structures usually belong to salts or metal-containing compounds which are not measured in this form in mass spectrometry. We ended up with 414,049 spectra with proper metadata annotation.

Quality assessment and filtering. When collecting data from a large number of publicly accessible repositories, it is important to consider that some, or even many, of the data might be incorrectly annotated. This issue arises because not all spectra originate from reference measurements; many spectra are annotated via spectral library searches or computational tools before being uploaded to public repositories. In the following analysis, we aim to remove compounds from the training data if there is uncertainty about their correct labeling.

One straightforward check is to count how many peaks, or how much intensity, can be explained by fragment ions whose molecular formulas are subsets of the parent molecule's formula. We use SIRIUS [9] to decompose all peaks in the spectra and remove any spectra from the dataset that explain less than 50% of the total intensity. To prevent a single large, intense peak from dominating the statistics, we apply a square root transformation to all peak intensities (here and in all subsequent filtering steps). We removed 135,190 spectra using this method. Although this number seems high, it only accounts for 926 unique molecular structures. Additionally, we found that the majority of the removed spectra did not contain a single peak that could be explained with a molecular formula subset of the parent molecule.

Next, we compared the spectra against the NIST20 [10] library – a commercially available, high-quality, manually curated spectral library. We found that one third of the structures are contained in both spectral libraries. For 14,600 spectra (38 unique structures), there was not a single peak (beyond the parent peak) shared with the NIST spectra of the same molecular structure. For an additional 9,290 spectra, the maximum cosine similarity with any NIST spectrum of the same structure was below 0.2, or there were fewer than 5 shared peaks. We removed these spectra; again, the number of removed structures was much lower, with only 126 structures removed.

Finally, we compared all spectra of the same structure within our dataset pairwise to identify outliers. We only considered structures with at least 4 spectra and removed 11,740 spectra for which the maximum cosine similarity to any other spectra of the same structure was lower than 0.2. We used such a conservative threshold to avoid removing spectra solely because they were measured at a very different collision energy than the other spectra.

In total, we are left with 231,104 spectra. A manual examination of a small random subset of the removed spectra found all of them at least "suspicious," with many clearly being wrongly annotated.

For each of these remaining spectra, we additionally provide fragment peak-molecular formula annotations in the form of fragmentation trees [12, 13, 14]. Each fragmentation tree annotates the

Table 2: **Summary of our data cleaning pipeline.** The rows in the table provide a report on the number of removed spectra, the number of spectra with modified metadata, and the number of spectra that were removed after each filtering step. The filtering steps were applied sequentially from top to bottom, as listed in the table. A horizontal rule separates the matchms filters from the additional filters that were applied beyond matchms.

Filter	Removed spectra	Changed metadata	Changed mass spectrum
add_parent_mass	0	699,187	0
add_precursor_formula	0	464,082	0
add_retention_index	0	706,484	0
add_retention_time	0	552,948	0
clean_adduct	0	8,102	0
clean_compound_name	0	104,201	0
correct_charge	0	322,993	0
derive_adduct_from_name	0	376,859	0
derive_annotation_from_compound_name	0	53,716	0
derive_formula_from_name	0	47,156	0
derive_formula_from_smiles	0	293,157	0
derive_inchi_from_smiles	0	42,470	0
derive_inchikey_from_inchi	0	413,405	0
derive_ionmode	0	137,815	0
derive_smiles_from_inchi	0	38,547	0
harmonize_instrument_types	0	294,479	0
harmonize_undefined_inchi	0	117,746	0
harmonize_undefined_inchikey	0	467,179	0
harmonize_undefined_smiles	0	114,075	0
normalize_intensities	0	0	513,299
remove_charged_molecules	4,776	0	0
remove_instrument_types	13,481	0	0
remove_noise_below_frequent_intensities	0	0	64,805
remove_not_ms2_spectra	628,478	0	0
repair_adduct_based_on_smiles	0	305,728	0
repair_inchi_inchikey_smiles	0	33,319	0
repair_not_matching_annotation	0	1,695	0
repair_smiles_of_salts	0	6,596	0
require_adduct_in_list	41,049	0	0
require_correct_ionmode	153,321	0	0
require_formula_match_parent_mass	152	0	0
require_matching_adduct_precursor_mz_parent_mass	3,726	0	0
require_minimum_number_of_peaks	1,357	0	0
require_number_of_peaks_below_maximum	880	0	0
require_parent_mass_match_smiles	8,694	0	0
require_precursor_mz	7,291	0	0
require_valid_annotation	22,036	0	0
store_relevant_metadata_only	0	449,721	0
make_charge_int	0	0	0
add_compound_name	0	0	0
interpret_pepmass	0	0	0
add_precursor_mz	0	0	0
require_matching_adduct_and_ionmode	0	0	0
Remove salts	3,819	0	0
< 50% explained intensity by molecular formula decomposition	135,190	0	0
No peaks shared with NIST spectrum	14,699	0	0
Cosine similarity against NIST < 0.2	9,290	0	0

fragment peaks of the corresponding MS/MS spectrum with sub-formulas of the parent's molecular formula. These fragmentation trees were computed using SIRIUS (version 6.0.4) with default parameter settings.

Subsetting spectra for the spectrum simulation challenge. Since instrument type and collision energy are required inputs for the spectrum simulation challenge, we consider only a subset of the MassSpecGym dataset for this challenge. We note that in this dataset, 98% percent of entries contain the [M+H]+ adduct. Therefore, we additionally removed the other 2% of entries with the [M+Na]+ ionization adduct for the simulation challenge. The boolean mask for this subset is stored in the dataset under the variable named `simulation_challenge`.

2.4 Data standardization

Spectrum standardization. We standardize all MS/MS spectra to have relative intensity values. Specifically, for each MS/MS spectrum, we divide each intensity value by the maximum intensity

within that spectrum. This normalization process ensures that the spectra are comparable regardless of their absolute intensity values, which can vary due to differing MS experimental conditions.

Instrument standardization. Instrument type descriptions were standardized based on the fixed vocabulary for instrument types of the MassBank record format (`https://github.com/MassBank/MassBank-web/blob/dev/Documentation/MassBankRecordFormat.md#2.4.2`) as a basis. For records originating from MassBank or having an instrument type description matching a MassBank notation (e.g., `LC-ESI-ITFT`), the instrument type was retained. For all other records, a mapping was manually generated from the unique values for instrument types in GNPS records. Explicitly specified instrument names were assigned to the code for their instrument types. Clearly nonsensical entries (e.g., `ESI-LC-ESI-IT`, which has two ionization mechanism tags) were mapped to the most plausible explanation (here, `LC-ESI-IT`). Some entries are imprecise and therefore ambiguous. In particular, `Orbitrap` may either correspond to `ITFT` or `QFT`. Finally, for the purpose of the final dataset, only the analyzer type is relevant to the type of spectrum observed, and details related to sample introduction (e.g. `LC`) are irrelevant to the measured data. Therefore, in the dataset, only the analyzer type is reported, where `QTOF` represents all instruments based on a quadrupole-time of flight analyzer and `Orbitrap` represents all instruments with an Orbitrap-type analyzer.

Collision energy standardization. The provided collision energies were standardized to appropriately handle normalization. Collision energy normalization is way of representing the fragmentation energy that accounts for the mass of the precursor ion. Assuming the precursor p is singly charged (which was the case for the spectra in our dataset, since we removed all spectra whose adducts were not [M+H]+ or [M+Na]+), the non-normalized collision energy $\mathrm{CE}(p)$ can be calculated from the normalized collision energy $\mathrm{NCE}(p)$ and the associated precursor m/z $\mathrm{m}(p)$ using Equation 3.

$$\mathrm{CE}(p) = \frac{\mathrm{m}(p) \times \mathrm{NCE}(p)}{500} \tag{3}$$

This transformation was applied to all collision energies that were identified as normalized. In cases of ambiguity, we assumed that the provided collision energy was not normalized.

Molecular structure standardization. Finally, the molecular structures of all data sets were standardized with respect to their corresponding SMILES strings in order to have a unified and unique representation of compound structures. SMILES were standardized using the PubChem standardization service accessed through the standardizeUtils python package (`https://github.com/boecker-lab/standardizeUtils`), which utilizes the PubChem Power User Gateway (PUG) [15]. To facilitate standardization of newly generated compounds, this is also directly possible from MassSpecGym. Compounds whose SMILES representation failed standardization were discarded.

2.5 Data splitting

To obtain the training-validation-test split, we apply an MCES-based clustering technique. First, we compute the matrix of pairwise MCES distances between all molecules in the dataset with unique PubChem-standardized SMILES strings. Here, MCES distances are computed using the threshold of 10 with the `always_stronger_bound` option disabled. Then, we use this matrix to perform agglomerative clustering with the minimum distance as the cluster linkage criterion. Clusters are linked together only if the minimum distance is lower than 10. Specifically, we use the following initialization of the agglomerative clustering from the scikit-learn Python package [17]: `AgglomerativeClustering(metric=‘‘precomputed’’, linkage=‘‘single’’,`

Table 2: **Composition of the dataset split with respect to the number of distinct spectra, molecules, and MCES-based clusters**. The subtable below "All metadata available" describes the simulation challenge subset.

	Num. spectra	Num. molecules	Num. clusters
Training	194,119 (84%)	25,046 (79%)	3,061 (41%)
Validation	19,429 (8%)	3,386 (11%)	2,221 (30%)
Testing	17,556 (8%)	3,170 (10%)	2,202 (29%)
All metadata available			
Training	101,573 (84%)	13,543 (74%)	2,628 (41%)
Validation	9,975 (8%)	2,445 (13%)	1,917 (30%)
Testing	10,159 (8%)	2,417 (13%)	1,907 (30%)

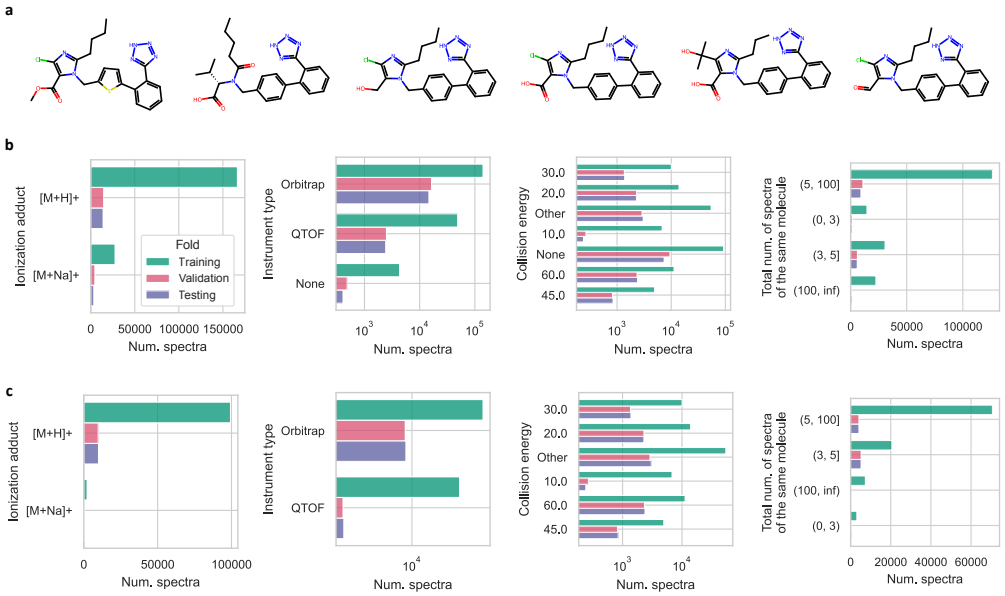

Figure 3: **MCES-based metadata-stratified data split results in a balanced composition of metadata across data folds. a,** Example of an MCES cluster of size six. **b,** Number of spectra in each fold in the dataset with respect to different metadata properties. **c,** Same as **b**, but for the subset of the dataset with no missing metadata. This subset is used for the spectrum simulation challenge.

`distance_threshold=10, n_clusters=None)`. Figure 3a shows an example of a cluster. We leverage the computed clusters as groups for the `StratifiedGroupKFold` splitting algorithm, assigning each molecule to one of the three folds. We use the concatenated string representations of ionization adducts, collision energies, instrument types, and the numbers of spectra per molecule as the stratification labels.

Since the majority (65%) of molecules formed a single cluster, the initial splitting resulted in a 74%-13%-13% composition of training-validation-test folds in terms of the number of dataset entries (i.e., spectra) and a 65%-18%-17% composition in terms of molecular structures. Such proportions are generally acceptable for the training-validation-test split. However, we found that the resultant split underrepresents molecules in the training dataset when considering a subset with all metadata available (55%-22%-22%), as required for the simulation challenge. To address this issue, we reassigned randomly sampled validation and testing clusters to the training fold to obtain a more balanced composition of the split with respect to molecular structures (74%-13%-13%). Table 2 shows the final proportions after the reassignment. Figure 3b and Figure 3c show the balanced proportions of metadata values across folds for the full dataset and the metadata-complete subset respectively. Figure 4 shows the balanced distribution of chemical classes across the entire dataset.

2.6 Auxiliary unlabeled datasets

Additionally, as a part of our MassSpecGym dataset, we provide curated unlabeled datasets of mass spectra and molecules. For the mass spectra, we provide the GeMS-A10 dataset, a deduplicated collection of 24 million high-quality mass spectra mined from the MassIVE GNPS repository [18]. As detailed in the main text, for the molecules, we provide three datasets of increasing size and decreasing relevance for mass spectrometry applications: a 1-million set [3] and a 4-million set [19] of biologically significant molecules, and all 118 million molecules (as of May 31, 2024) from PubChem [20].

We use the sets of molecules to construct retrieval candidates for the challenges of molecule retrieval and spectrum simulation. The procedure is outlined in Algorithm 1. The algorithm iteratively samples candidates from given molecular databases that are similar to the query molecule until the maximum number of candidates is reached. The similarity is measured either by similar mass (up

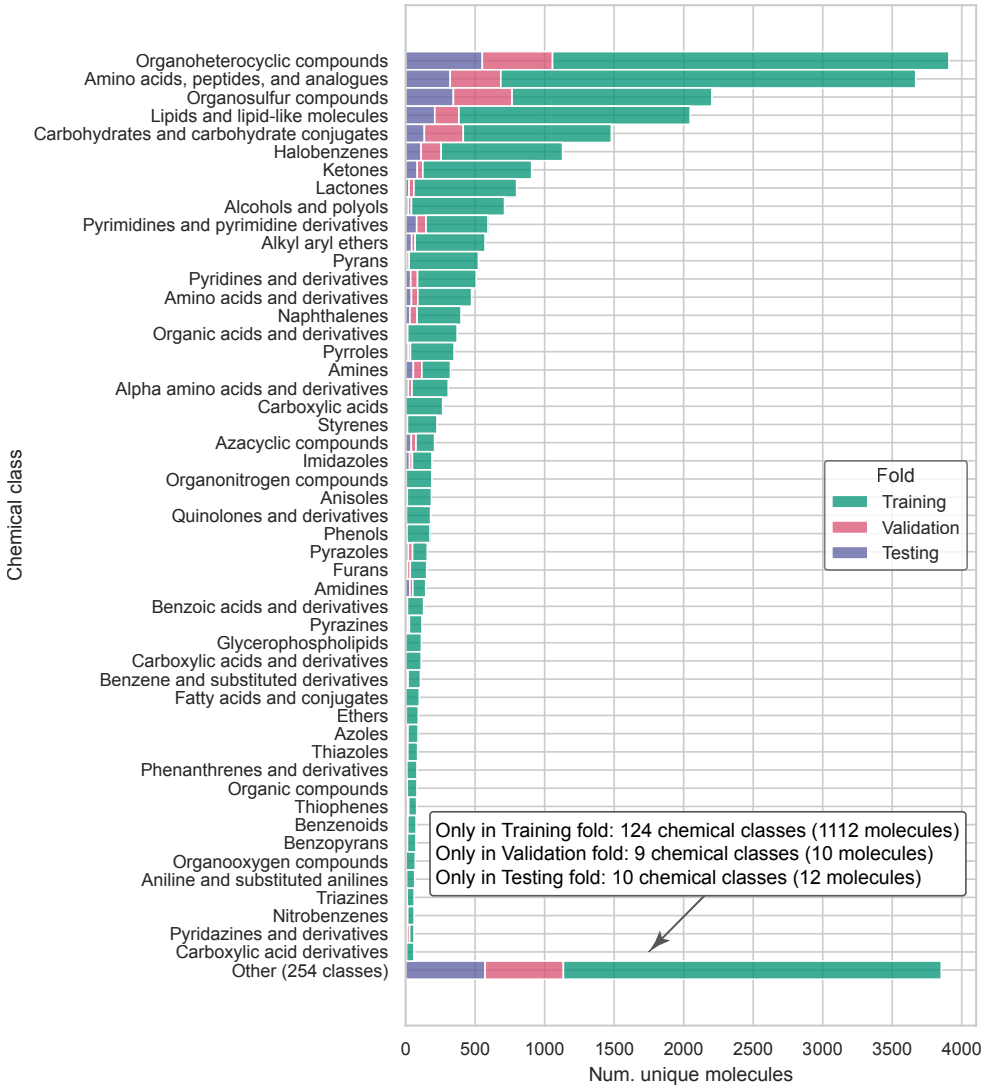

Figure 4: **MCES-based metadata-stratified data splitting results in a balanced distribution of chemical classes across data folds.** The figure presents a histogram of the 50 most common chemical classes according to ClassyFire [16], found in MassSpecGym, with a separate bar for all less common classes, labeled as "Other (254 classes)". The box with an arrow pointing to this bar shows the number of classes uniquely present in individual folds, along with the number of underlying molecules.

to experimental error) for the standard challenges, or identical formula for the bonus challenges. In practice, we use the three chemical databases mentioned above as $\mathcal{D}$ and the maximum number of candidates $|C| = 256$. The statistics and examples of retrieval candidates are visualized in Figure 5.

When developing new methods that leverage unlabeled data, we anticipate that users will rely solely on the following two datasets: the GeMS-A10 dataset of unlabeled MS/MS spectra and a refined subset of the unlabeled 4 million-molecule dataset. The molecular dataset has been refined by excluding any molecules with an MCES distance of less than two from any molecule in the test fold of MassSpecGym. This refinement is intended to reduce the potential for data leakage, particularly when used in the context of the *de novo* generation challenge.

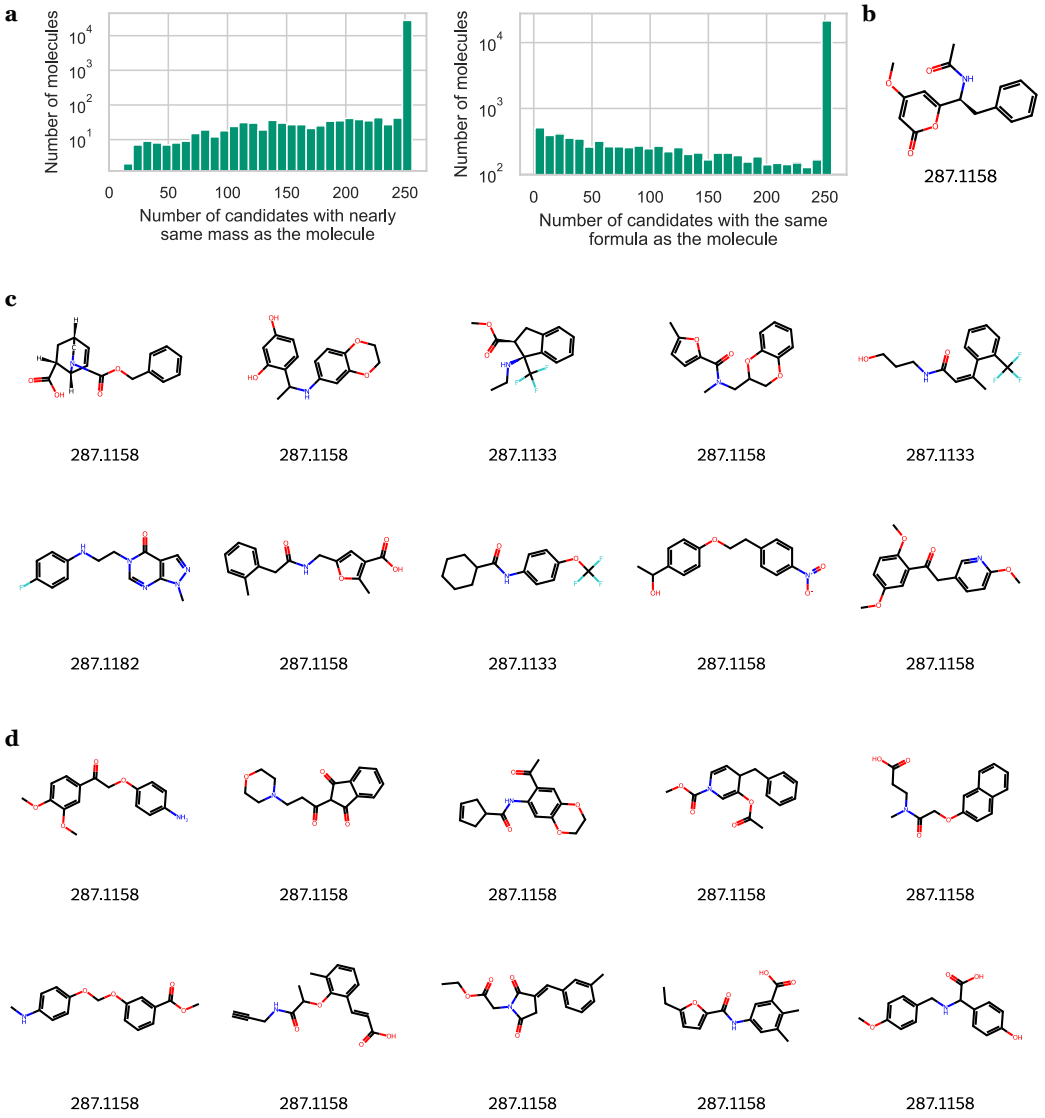

Figure 5: **Overview of molecule retrieval candidates. a,** The distributions of the number of retrieval candidates per molecule in the MassSpecGym dataset. The left histogram corresponds to candidates for the main molecule retrieval challenge, where the candidates are selected to have a mass similar to the query molecule, and the right histogram corresponds to the chemical formula bonus challenge, where candidates are restricted to having the same formula as the query molecule. Most of the molecules (87% and 66% respectively) have the predefined maximum number of 256 candidates. **b,** Example of a molecule having 256 both mass-based and formula-based candidates. **c,** Ten randomly sampled mass-based candidates for the molecule shown in figure **b**. **d,** Ten randomly sampled formula-based candidates for the molecule shown in figure **b**. The numbers below the molecules show their mass (in Daltons).

2.7 Dataset limitations

MassSpecGym aims to make machine learning applied to MS/MS spectra accessible to the machine learning community and rigorously standardized. To achieve this, we had to make certain simplifications that inherently limit MassSpecGym. For example, we focused exclusively on MS/MS spectra acquired in positive ionization mode, retained only spectra with the most common ionization adducts and with singly-charged ions, and sought to eliminate noisy spectra. However, this approach does not

Algorithm 1 Construction of retrieval candidates

Require: Query molecule q, kind of candidates $kind \in \{\texttt{mass}, \texttt{formula}\}$, list of chemical databases with increasing significance $\mathcal{D}$ to sample candidates from, maximum number of candidates N
Ensure: Candidate set $\mathcal{C}$

```
C ← {q}
if |C| = N then
    return C
end if
for D ∈ D do
    if kind = mass then
        ε ← mass(q) × 10 × 10^−6                          ▷ 10 parts per million (ppm)
        C ← C ∪ {c ∈ D | |mass(c) − mass(q)| < ε ∧ inchi2d(c) ≠ inchi2d(q)}
    else if kind = formula then
        C ← C ∪ {c ∈ D | formula(c) = formula(q) ∧ inchi2d(c) ≠ inchi2d(q)}
    end if
end for
C ← C[: N]                                    ▷ Get first N elements added to C
return C
```

fully reflect real-world scenarios where a significant portion of spectra may be instrument noise or acquired in different settings. Additionally, we were unable to adequately address certain types of noise, such as isotope signals in MS/MS spectra or chimeric spectra representing multiple molecules, due to the lack of appropriate metadata in public spectral libraries. Our dataset is also combined from highly heterogeneous and imbalanced spectral libraries, which may pose challenges when training machine learning models. Finally, our work is solely focused on MS/MS spectra, not considering other types of spectra, such as electron ionization (EI) or nuclear magnetic resonance (NMR) spectra.

3 Evaluation of baselines

In this section, we present a detailed description of the architectures, loss functions, and hyperparameter configurations for each model. Unless stated otherwise, we train the models using all possible combinations of hyperparameters and select the best model for testing based on the minimum validation loss. To prevent overfitting, training is stopped if there is no improvement in validation loss over the last three validation epochs. Unless specified, we train the models on four AMD MI250X GPUs (8 PyTorch devices) using the distributed data parallel (DDP) mode.

3.1 De novo molecule generation

Random chemical generation. The goal of the challenge is to assess the capabilities of predictive models to extract information about molecules from their mass spectra. A potential failure mode for these models is to ignore the input mass spectra and generate molecules solely from prior chemical knowledge. To address this risk, we implement a baseline that generates molecular structures using only prior chemical knowledge.

In detail, the baseline consists of combinatorial and graph algorithms. It begins by identifying a chemical formula from the training data with the closest molecular weight. Then, the procedure iterates over plausible combinations of atom valence assignments that satisfy the selected chemical formula and can form a feasible combination of covalent and coordinate bonds in a molecule. Finally, through random graph traversal, we generate a connected molecular structure that respects the assigned atomic properties (valence and charge). Additionally, the baseline samples the edges in molecules based on the edge distribution observed in the training data structures.

SMILES Transformer. The SMILES Transformer model uses a standard encoder-decoder architecture [21] with post-norm and is trained using standard teacher forcing method. The input is given by 2D continuous tokens capturing m/z and intensity values of the spectrum peaks. Additionally, we add a token representing precursor m/z with a dummy intensity equal to 1.1. We linearly project the tokens to have dimensions corresponding to the hidden dimension of the model. The output SMILES

Table 3: **Hyperparameter grid explored for SMILES Transformer and SELFIES Transformer on the *de novo* generation challenge.** The optimal values, leading to the minimum validation loss, are highlighted in bold. For the bonus chemical formulae challenge, we use the same hyperparameters.

Hyperparameter	Values (SMILES Transformer)	Values (SELFIES Transformer)
Learning rate	$\mathbf{3 \cdot 10^{-4}}$, $1 \cdot 10^{-4}$, $5 \cdot 10^{-5}$	$\mathbf{3 \cdot 10^{-4}}$, $1 \cdot 10^{-4}$, $5 \cdot 10^{-5}$
Batch size (per GPU)	64, **128**	64, **128**
k predictions	**10**	**10**
Transformer hidden dimensionality	**256**, 512	**256**, 512
Number of attention heads	**4**, 8	4, **8**
Number of encoder layers	**3**, 6	3, **6**
Sampling temperature	0.8, **1.0**, 1.2	**1.0**

Table 4: **Hyperparameter grid explored for Fingerprint FFN on the molecule retrieval challenge.** The optimal values, leading to the minimum validation loss, are highlighted in bold. For the bonus chemical formulae challenge, the optimal hyperparameters remain the same, except for an increased batch size of 64.

Hyperparameter	Values
Learning rate	$3 \cdot 10^{-4}$, $\mathbf{1 \cdot 10^{-3}}$
Batch size (per GPU)	**16**, 64, 128
Hidden dimensionality	**128**, 1024
Number of layers	3, **7**
Dropout	0.0, **0.1**

is postprocessed using a byte-pair encoder [22] to increase the chance of generating valid molecules. The byte-pair encoder is trained on the 4M set of molecules discussed in Section 2.6. We use greedy decoding by sampling each token from the predicted distribution over the softmax vocabulary with a given temperature to predict k samples for each input. For early stopping, we monitor the validation loss once per epoch. The training of the best model took 24 epochs (1 hour) in total. Hyperparameters used in the model are listed in Table 3.

For the bonus chemical formulae challenge, we slightly modify the model to condition it on the chemical formula of a target molecule. Specifically, we feed the formula into the encoder alongside the MS/MS spectrum using an additional feed-forward network. This network maps the vectorized chemical formula (where each element represents the count of a specific chemical element) to the transformer's hidden dimensionality. The output from this network is then added to the embedding of each token before it enters the transformer layers. We use a single hidden layer with a dimensionality matching the number of considered chemical elements, which is 118.

SELFIES Transformer. The SELFIES Transformer model follows the same implementation as the SMILES Transformer, with the exception of using SELFIES molecular string representations [23] instead of SMILES. Additionally, we do not use byte-pair encoding for this model, as the SELFIES grammar is specifically designed for generating molecular graphs that are both syntactically and semantically valid.

The optimal hyperparameters for SELFIES Transformer are provided in Table 3, and the training of the best model took 25 epochs (1 hour) in total.

3.2 Molecule retrieval

Fingerprint FFN. The Fingerprint FFN baseline employs a simple feedforward neural network to predict Morgan fingerprints of molecules from binned MS/MS spectra. Specifically, we convert an input spectrum using binning with a maximum m/z value of 1005 and a bin width of 1 m/z. The network outputs a 4096-dimensional vector, trained to approximate the true Morgan fingerprint (with a radius of 2) of the underlying molecule, using cosine similarity as the loss function. Once the fingerprint is predicted for a spectrum, we sort the corresponding candidate list by cosine similarity

to obtain the final top-k predictions. For the chemical formulae bonus challenge, we use a different set of candidates for each spectrum, pruned to include only molecules with the same formula.

We experiment with various standard feedforward neural network hyperparameters. The grid of values explored in this work is provided in Table 4. We monitor the validation loss twice per epoch. The training of the best model took 4 epochs (6.5 hours) for the standard challenge and 5 epochs (7 hours) for the bonus challenge.

Table 5: **Hyperparameter grid explored for DeepSets on the molecule retrieval challenge.** The optimal values, leading to the minimum validation loss, are highlighted in bold. For the bonus chemical formulae challenge, the optimal hyperparameters for the architecture remain the same while the training parameters are a learning of $1 \cdot 10^{-3}$ and a batch size of 128.

Hyperparameter	Values
Learning rate	$\mathbf{3 \cdot 10^{-4}}$, $1 \cdot 10^{-3}$
Batch size (per GPU)	$\mathbf{16}$, 64, 128
Hidden dimensionality	$\mathbf{128}$, 1024
Number of layers (per MLP)	2, $\mathbf{4}$
Dropout	0.0, $\mathbf{0.1}$

DeepSets. Since the binning of spectra has the drawback of producing sparse input vectors and ambiguously rounding signals at the edges of m/z bins, we implement a DeepSets baseline [24] that treats MS/MS spectra as sets of two-dimensional elements: m/z and intensity values of individual signals. The output of DeepSets is a 4096-dimensional Morgan fingerprint, similar to Fingerprint FFN. The entire model is represented as:

$$\hat{\mathbf{y}} = \rho\left(\sum_{i=1}^{n} \phi(\mathbf{m}_i \parallel \mathbf{i}_i)\right), \tag{4}$$

where $\hat{\mathbf{y}}$ is the predicted fingerprint, $\phi : \mathbb{R}^2 \to \mathbb{R}^d$ and $\rho : \mathbb{R}^d \to \mathbb{R}^{4096}$ are feed-forward neural networks, d is the hidden dimensionality, n is the number of signals in the input MS/MS spectrum, $\mathbf{m}_i$ and $\mathbf{i}_i$ are m/z and intensity values of the i-th signal respectively. $\parallel$ denotes concatenation.

Similar to the Fingerprint FFN baseline, we experiment with different hyperparameter setups (Table 5), monitoring validation performance twice per epoch for early stopping. The training of the best model took 3 epochs (10.5 hours) for the standard challenge and 1 epoch (5 hours) for the bonus challenge.

DeepSets + Fourier features. To enhance the representation of m/z values in the DeepSets architecture, we process them using the Fourier features technique [25]. Specifically, we pre-process each m/z value with a set of sine and cosine functions with 6,000 pre-defined wavelengths, following [18]. These wavelengths decompose each m/z value into 6,000 input features, capturing its both integer and decimal components with greater sensitivity than the single value. This allows the model to be, for example, more sensitive to small differences in m/z values, which are particularly relevant when working with high-accuracy mass spectrometers. The overall DeepSets architecture is then modified as follows:

$$\hat{\mathbf{y}} = \rho\left(\sum_{i=1}^{n} \phi\left(\mathbf{W}_1 \textsc{Fourier}(\mathbf{m}_i) + \mathbf{b}_1 \parallel \mathbf{W}_2 \mathbf{i}_i + \mathbf{b}_2\right)\right), \tag{5}$$

where $\textsc{Fourier} : \mathbb{R} \to \mathbb{R}^{6000}$ represents the Fourier features, and $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times 6000}$, $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times 1}$, $\mathbf{b}_1 \in \mathbb{R}^{d_1}$, and $\mathbf{b}_2 \in \mathbb{R}^{d_2}$ are learnable parameters. Here, $d_1 = \lceil 0.8d \rceil$, $d_2 = d - d_1$, and ϕ takes d inputs instead of 2.

For the DeepSets + Fourier features model, we use the final hyperparameters from the DeepSets model.

Table 6: **Hyperparameter values explored for MIST for the molecule retrieval challenge.** No grid search was performed due to the computational cost of training MIST. Instead, 5 combinations of hyperparameter values were selected for training. The optimal values, leading to the minimum validation loss, are highlighted in bold.

Hyperparameter	Values
Learning rate	0.0001, **0.0003**
Scheduler	True, **False**
Weight decay	1e-6, **1e-7**
Learning rate decay	**0.9**, 0.995
Dropout	0, **0.3**
Hidden size	**512**
Number of layers	**3**, 5
Number of unfolding layers	**5**
Unfolding loss weight	**0.1**
Magma loss weight	2, **4**
Probability noised spectrum	**0.5**
Probability remove peak	**0.5**
Probability scale peak	**0.1**
Batch size	**128**

MIST. The Metabolite Inference with Spectrum Transformers (MIST) [26] tool is employed as the state-of-the-art baseline for the molecule retrieval challenge. MIST utilizes a deep learning approach to annotate mass spectra with chemical structures. This tool integrates domain knowledge into its architecture by representing peaks with their associated chemical formulae, which are then processed by the specialized "chemical formula transformer". Furthermore, MIST predicts low-resolution fingerprints, which are subsequently refined to achieve full resolution. By predicting these fingerprints and matching them against a database of candidate structures, MIST effectively annotates MS/MS spectra.

Before training, the dataset is processed to meet the requirements of MIST, including converting spectra to `.ms` files, subformula labeling, and MAGMa [27] substructure annotation. MIST v2.0.0 is trained using the hyperparameters specified in Table 6. No simulated spectra are used during training. All relevant code can be found on `https://github.com/Janne98/mist/tree/mass-spec-gym`.

We train MIST on a node equipped with two AMD Epyc 7452 Zen2 CPUs (totaling 64 cores) and a single NVIDIA Ampere A100 GPU. The training process utilizes early stopping with a patience of 20 epochs. The validation loss is computed after each epoch. Convergence is achieved after 32 epochs, which takes approximately 5 hours. The final model weights and the processed dataset can be found on `https://zenodo.org/records/11580401`

3.3 Spectrum simulation

FFN Fingerprint. The FFN Fingerprint model is based on an implementation of the NEIMS model [28] from [29]. Let G be the input molecule, and $p_{\text{FP}}(G) \in \mathbb{R}^n$ be its molecular fingerprint featurization. In practice we use a combination of three molecular fingerprints: the ECFP4 (Morgan) fingerprint [1], the rdkit fingerprint [30], and the MACCS fingerprint [31]. Let $z \in \mathbb{R}^m$ be a vector representation of metadata for the spectrum prediction (collision energy, instrument type, and precursor adduct). Let $f_{\text{FP}} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ be a neural network that maps from the fingerprint to the binned spectrum, where d is the number of bins. The architecture of f_{FP} is an MLP with skip-connections. To predict the output spectrum, our model uses gated bidirectional spectrum prediction (refer to [28] for full details). The loss function is cosine distance, although the intensities of the binned target spectrum are subjected to a square-root transform before distance calculation.

Key hyperparameters for the FP model were optimized using a random sweep with a budget of 50 samples. The results of the sweep are summarized in Table 7. Performance was measured on the simulation split validation set (roughly 10% of all simulation spectra), and the configuration with the highest score was selected for testing. All models were trained for 100 epochs. The learning rate followed a linear decay schedule following a warmup of 1000 steps. The "Precursor m/z offset"

Table 7: **FFN Fingerprint model hyperparameter sweep configuration and optimal values for the spectrum simulation challenge.**

Hyperparameter	Possible Values	Optimal Value
Learning rate	[1e-4, 1e-3]	1.563e-4
Weight decay	{0, 1e-7, 1e-6}	1e-7
Learning rate decay	[0.7, 1.0]	0.9866
Batch size	{32, 64, 128, 256}	256
MLP Dropout	{0, 0.1, 0.2, 0.3, 0.4, 0.5}	0
MLP Hidden size	{256, 512, 1024}	1024
MLP Number of layers	{1, 2, 3, 4, 5}	5
Precursor m/z offset	{5, 50, 500}	5

Table 8: **Molecular graph node and edge features for the GNN model and the FraGNNet model.**

Feature	Possible Values
Atom Type (Element)	{C, O, N, P, S, F, Cl, Br, I, Se, Si}
Atom Degree	$\{0, \ldots, 10\}$
Atom Orbital Hybridization	{SP, SP2, SP3, SP3D, SP3D2}
Atom Formal Charge	$\{-2, \ldots, +2\}$
Atom Radical State	$\{0, \ldots, 4\}$
Atom Ring Membership	{True, False}
Atom Aromatic	{True, False}
Atom Mass	$\mathbb{R}^+$
Atom Chirality	{Unspecified, Tetrahedral CW, Tetrahedral CCW}
Bond Degree	{Single, Double, Triple, Aromatic}

parameter refers to the offset used for bidirectional prediction, corresponding to the τ parameter in Equations 4 and 5 of [28].

GNN Model. The GNN model is based on an implementation from [29]. Given the molecular graph $G = (V, E)$, a set of node and edge features are $X_V = \{X_v\}_{v \in V}$ and $X_E = \{X_e\}_{e \in E}$ are extracted (see Table 8). These features are passed to a GNN $g_{\text{GNN}}(G, X_V, X_E)$ which outputs a set of node embeddings $H_V = \{h_v \in \mathbb{R}^l\}_{v \in V}$. The GNN architecture is Graph Isomorphism Network with Edge Features [32], as implemented in the Pytorch Geometric library [33]. Average pooling is used to produce a graph-level embedding $h_G \in \mathbb{R}^l$ from the individual atom-level embeddings H_V. The molecule embedding is then concatenated with the metadata vector $z \in \mathbb{R}^m$ and passed to another neural network $f_{\text{GNN}}(h_G, z)$ that makes a spectrum prediction. The architecture of f_{GNN} is identical to that of f_{FP}, albeit with different hyperparameter configurations and input dimensions. The loss function and intensity transformations are consistent with the FFN Fingerprint model.

As was the case with the FFN Fingerprint model, a hyperparameter sweep with a budget of 50 samples was used to select hyperparameters for the GNN model. The results of this sweep are summarized in Table 9.

FraGNNet Model. The FraGNNet model [29] simulates a spectrum by modelling a distribution over fragments of the input molecule G and then mapping this distribution to a mass spectrum; for full details, refer to [29]. A recursive fragmentation algorithm $p_{\text{FRAG}}(G)$ is used to pre-process G into a fragmentation DAG $G_D = (V_D, E_D)$ that represents a hierarchy of plausible molecular fragments. Each node s of V_D represents a fragment (connected subgraph) of the original molecular graph G. A GNN $g^1_{\text{FRAG}}(G, X_V, X_E)$, with identical architecture to g_{GNN} (excluding hyperparameter choices), is used to generate atom embeddings H_V. The information from the DAG G_D and the atom embeddings H_V is jointly processed by a second GNN $g^2_{\text{FRAG}}(G_D, H_V)$, which produces embeddings $H_S = \{h_s \in \mathbb{R}^k\}_{s \in G_V}$. An MLP $f_{\text{FRAG}} : \mathbb{R}^k \to \mathbb{R}$ is applied to each fragment embedding to predict a distribution over fragments. The mass spectrum (which is a distribution over masses) can be inferred from the fragment distribution by aggregating probabilities for fragments that share the same chemical formula and using the masses of these formulae (which are known) to map the probabilities

Table 9: **GNN hyperparameter sweep configuration and optimal values for the spectrum simulation challenge.**

Hyperparameter	Possible Values	Optimal Value
Learning rate	[1e-4, 1e-3]	3.755e-4
Weight decay	{0, 1e-7, 1e-6}	1e-6
Learning rate decay	[0.7, 1.0]	0.8523
Batch size	{32, 64, 128, 256}	128
MLP Dropout	{0, 0.1, 0.2, 0.3}	0
MLP Hidden size	{256, 512, 1024}	512
MLP Number of layers	{2, 3, 4, 5}	4
Precursor m/z offset	{5, 50, 500}	5
GNN Normalization	{none, batch, layer, graph}	batch
GNN Dropout	{0, 0.1, 0.2, 0.3}	0.2
GNN Hidden size	{64, 128, 256, 512}	256
GNN Number of layers	{2, 3, 4, 5, 6}	4

Table 10: **FraGNNet hyperparameter values for the spectrum simulation challenge.**

Hyperparameter	Value
Learning rate	3.033e-4
Weight decay	0.01
Batch size	128
MLP Dropout	0.2
MLP Hidden size	64
MLP Number of layers	2
Mol GNN Normalization	batch
Mol GNN Dropout	0.2
Mol GNN Hidden size	256
Mol GNN Number of layers	4
Frag GNN Normalization	graph
Frag GNN Dropout	0.1
Frag GNN Hidden size	256
Frag GNN Number of layers	4

to a mass distribution. The model is trained by minimizing the cross-entropy between the predicted distribution and the ground-truth spectrum. Although FraGNNet can model peak locations with extremely high precision, during evaluation we bin the predicted spectrum at 0.01 Da to allow for fair comparison with the other baseline models.

The hyperparameters for the FraGNNet model were not optimized using a sweep, since the initial values significantly outperformed the other two simulation models. The key hyperparameter selections are summarized in Table 10. Unlike the other baselines, a learning rate schedule was not applied during training.

3.3.1 Jensen-Shannon Similarity Metric

The Jensen-Shannon similarity metric is an alternate way of representing Jensen-Shannon divergence between the true mass spectrum x and the predicted mass spectrum $\hat{x}$, both of which can be interpreted as discrete probability distributions. It is defined using the following equation:

$$JSS(x, \hat{x}) = 1 - \frac{JSD(x, \hat{x})}{\log(2)} \tag{6}$$

Note that JSD denotes the Jensen-Shannon divergence calculated with the natural logarithm $\log$. JSD ranges from $\log(2)$ (when x and $\hat{x}$ share no support) to 0 (when $x = \hat{x}$); JSS ranges from 0 to 1.

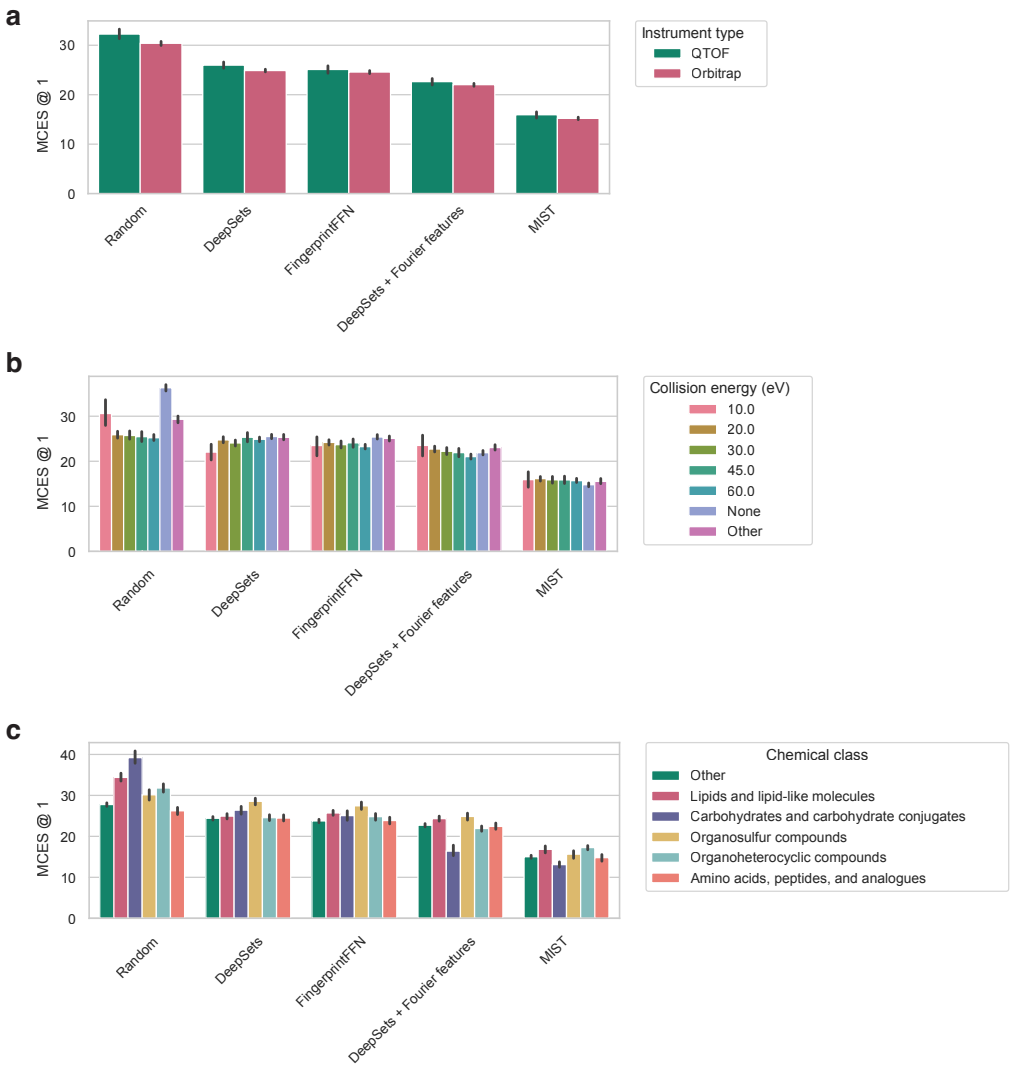

Figure 6: **Test performance of the MIST model on the molecule retrieval challenge, stratified by various metadata. a,** Performance stratified by instrument types. **b,** Performance stratified by collision energies. **c,** Performance stratified by the top 5 most common chemical classes. Error bars represent 99.9% confidence intervals, calculated through bootstrapping with 20,000 resamples.

3.4 Extended results

In addition to the overall metrics averaged across the entire test set, as presented in the main text, we evaluate the MCES @ 1 metric on the molecule retrieval challenge across different metadata subsets. Specifically, we stratify the results of all baselines by instrument types, collision energies, and chemical classes (Figure 6). Overall, all methods perform better on MS/MS spectra from Orbitrap instruments, which generally produce higher-quality spectra. Regarding collision energies and chemical classes, machine learning methods perform relatively consistently across various subsets, with the exception of DeepSets + Fourier features, which excels on carbohydrates and carbohydrate conjugates. Interestingly, this chemical class exhibits the highest variability in performance across methods, while classes such as organosulfur compounds result in more limited variability.

4 Background on mass spectrometry

This section introduces mass spectrometry to the broader machine learning community. Based on [34], we first define the concept of molecular structures (Section 4.1) and then outline the basic principles of tandem mass spectrometry (Section 4.2).

4.1 Small molecules

An **atom** is the fundamental building block of matter. Each atom is formed of **protons** and **neutrons** comprising a nucleus, as well as **electrons** orbiting around the nucleus. The number of protons defines a **chemical element** of an atom. For example, H (hydrogen) atom has 1 proton, C (carbon) atom has 6 protons, Br (bromine) atom has 35 protons, and so on. A **molecule** is a compound made up of two or more atoms that are chemically **bonded** together by sharing electrons. The more pairs of electrons they share the stronger the bond between atoms. **Molecular structure** is typically understood as a labeled undirected graph, where nodes represent atoms and are labeled by the corresponding chemical elements along with their spatial coordinates. Edges represent bonds and are labeled by the number of shared electron pairs.

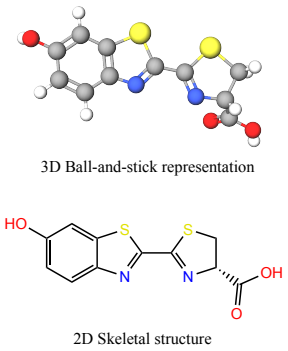

Figure 7: **Example molecular structure of *firefly luciferin* – a compound responsible for the characteristic yellow light emission from many firefly species.**

Molecular representations. Molecules containing carbon-hydrogen bonds are named **organic compounds**. Since they constitute the majority of known chemicals, it is convenient to represent them using simplified planar graph representations - **skeletal structures**. The example is shown in fig. 7. In a skeletal structure, nodes are implicitly associated with carbon atoms, and hydrogens adjacent to carbons are omitted. Noteworthy, this adaptation does not affect the expressivity of the representation because hydrogens can be unambiguously filled based on the bonding capacity of each element given by its composition. A spatial arrangement of atoms is encoded in special **stereochemical** types of bonds visualized as either dashed or solid triangles representing two opposite directions orthogonal to a molecular plane. Molecules differing only in stereochemical bonds are referred to as **stereoisomers**.

In order to simplify computer storage and processing, molecular structures are commonly encoded as strings. The three most widely used variants are **SMILES** (Simplified molecular-input line-entry system), **InChI** (International Chemical Identifier), and **InChIKey**. Although both SMILES and InChI serve the purpose of uniquely identifying a molecule as a sequence of characters, SMILES are more human-readable and simpler but do not undergo a unified standard. In contrast, InChI strings have more complex yet standardized grammar. To give an example, SMILES string of *firefly luciferin*, depicted in Figure 7, is C1[C@@H](N=C(S1)C2=NC3=C(S2)C=C(C=C3)O)C(=O)O, while its InChI string is InChI=1S/C11H8N2O3S2/c14-5-1-2-6-8(3-5)18-10(12-6)9-13-7(4-17-9)11(15)16/h1-3,7,14H,4H2,(H,15,16)/t7-/m1/s1. InChIKey representations are fixed-size hashes (e.g., IWJYWBVPCGUPLO-BFUDMSGGSA-N) derived from InChI strings, which are convenient, for instance, to perform searches of molecules in large databases. First 14 characters of the InChI key encode the connectivity information and are often referred to as **2D InChI keys**. Another compact coarse-grained representation of a molecule is its **chemical formula**, which represents the histogram of chemical elements within a molecule. Accordingly, the chemical formula of *firefly luciferin* is $C_{11}H_8N_2O_3S_2$.

The comparison of molecular structures is commonly conducted by utilizing their **fingerprints**, which are fixed-size binary vectors. A basic example of a molecular fingerprint is the Molecular ACCess System (MACCS) [35], where each of its 166 bits represents the presence or absence of a specific predefined substructure. The most widely-adopted family of fingerprints is Morgan fingerprints [1], which are fixed-size hashes encoding the local neighborhoods of molecular atoms. To compare two molecules, the most common approach is to generate the corresponding fingerprints and compute their Tanimoto similarity. The Tanimoto similarity is defined as a ratio of the number of shared positive bits to the total number of unique positive bits present in both fingerprints.

Molecular properties. **Molecular mass** is a central notion for mass spectrometry. It is characterized as a sum of **atomic masses** constituting the molecule and is usually measured in **Da** (Daltons). Single Dalton is defined as $\frac{1}{12}$ of the mass of ^{12}C (carbon atom containing 6 protons and 6 neutrons). Importantly, a molecule is roughly termed as a "**small molecule**" if its mass is less than 1000 Da. As a consequence of the definition of a Dalton and the significantly smaller mass of electrons compared to protons and neutrons, one nuclear particle has a mass approximately equal to 1 Da. Specifically, a proton has a mass of approximately 1.007 Da, a neutron has a mass of approximately 1.009 Da, and an electron has a mass of approximately 0.0005 Da[1]. While the number of protons in the atom of a chemical element is given by the definition, the notation ^{12}C is used to explicitly express the number of neutrons and protons.

Atoms having the same number of protons but different numbers of neutrons are referred to as **isotopes** of a chemical element. For instance, Cl (chlorine) element has two stable[2] isotopes: ^{35}Cl having a mass of 34.96885269(4) Da and ^{37}Cl having a mass of 36.96590258(6) Da. Naturally, ^{35}Cl occurs in roughly 76% of cases and ^{37}Cl occurs in remaining 24% of cases. Another element S (sulfur) has four stable isotopes: ^{32}S, ^{33}S, ^{34}S, and ^{36}S with natural abundances 94.99%, 0.75%, 4.25%, and 0.01% respectively. In contrast, F has only a single stable isotope ^{19}F. The mass of the most abundant isotope is often termed as a **monoisotopic mass** of an element.

Another molecular property essential for the domain of mass spectrometry is a **molecular charge**. A molecule is defined as **negatively charged**, **neutral**, or **positively charged** if its atoms in total have more, equal, or fewer electrons than protons respectively. A charged molecule is termed **ion** and its charge is often expressed as an integer indicating the difference between the number of protons and electrons. The sign of such an integer can be often found in different notations and depictions of a molecule. For example, a positively-charged ion of a molecule M can be denoted as M+. **Ionization adducts** are formed when a molecule binds with an ion. Common examples include the [M+H]+ adduct, where a molecule M gains a proton (H+), and the [M+Na]+ adduct, where a molecule M gains a sodium ion (Na+).

4.2 Tandem mass spectrometry

The identification of molecules present in a sample is a fundamental task in various fields of biology and environmental science. To achieve this, Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) is the most widely employed technique. This method constitutes an intricate pipeline composed of liquid chromatography and several stages of mass spectrometry (i.e., tandem mass spectrometry) to separate, elucidate, and quantify compounds in complex mixtures. Nevertheless, interpreting the output data from LC-MS/MS presents a significant challenge, as information about the molecules is only available in terms of their masses or the masses of their fragments. In the following section, we introduce tandem mass spectrometry. However, we do not discuss liquid chromatography, as it is outside the scope of this work.

Acquisition of mass spectra. Mass spectrometry (MS) plays a critical role in the workflow of LC-MS/MS, enabling the determination of the molecular mass of a compound. The fundamental principle of MS involves ionizing a sample to create charged molecules or fragments that are subsequently separated based on their mass-to-charge ratio (**m/z**) and detected using a mass analyzer. It is important to note that MS instruments are only capable of measuring m/z values and not the mass or charge[3] of the ions individually. The resulting **mass spectrum** is a collection of two-dimensional points, with the first dimension representing m/z values and the second dimension corresponding to their respective abundances (i.e. the **intensities** of the detected signals).

The ionization process in MS can occur through a variety of methods, including electron impact ionization (EI), electrospray ionization (ESI), and matrix-assisted laser desorption ionization (MALDI). Regardless of the method used, the result is the formation of ions with different m/z ratios, which are then separated by the mass analyzer. ESI is the most convenient and common method to couple with LC. During the electrospray ionization, the liquid sample is sprayed through a small needle

[1] Although the mass of an atom might be expected to equal the sum of the masses of its particles, it is always slightly less (with the exception of the hydrogen atom). This phenomenon is caused by the nuclear binding energy and is termed the mass defect of the nucleus.

[2] Isotope is stable if it does not decay into other elements on geologic timescales.

[3] Although, advanced mass spectrometry instruments report charges.

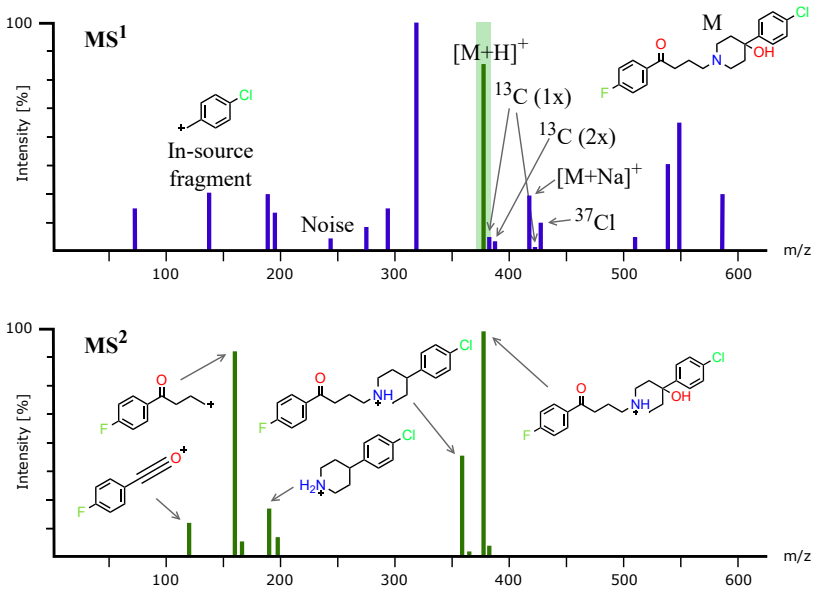

Figure 8: **Examples of MS^1 and MS/MS (i.e., MS^2) spectra. Top,** An example of an MS^1 spectrum, featuring haloperidol (designated as "M"), with a mass of 375.14 Da. Several peaks are labeled, while others may correspond to different compounds. Isotopes are denoted by ^{13}C (1x) and ^{13}C (2x), and ^{37}Cl, indicating ions ([M+H]$^+$ or [M+Na]$^+$) containing corresponding isotopes. Notice, that the spectrum is simplified for visualization purposes, and one can frequently observe more intricate isotopic patterns and adduct species. **Bottom,** MS/MS (i.e., MS^2) spectrum acquired by fragmenting protonated haloperidol from MS^1 belonging to the isolation window highlighted in green.

that has a high voltage applied to it. The high voltage causes the liquid to form tiny droplets, and as these droplets move through the air, they pick up an electrical charge. These droplets will either be positively or negatively charged depending on the polarity of the applied voltage. Further, the charged droplets continue to break apart and re-form until they eventually become individual ions. During the ionization process, ions form **adducts**, which are clusters of molecules that stick together due to electrostatic interactions. For example, a molecule in the sample may pick up a positively charged droplet ion such as a proton, Na (sodium), or K (potassium). As a consequence, the resulting ion measured in a mass spectrum has a higher mass-to-charge ratio than the original molecule. Such adduct species of molecule M are then denoted as [M+H]+, [M+Na]+, or [M+K]+ respectively.

It is important to note that the understanding of the sample introduced to MS system should not be limited to a set of mutually exclusive molecules. Instead, it should be understood as a complex mixture of compounds with millions of duplicates for each molecular structure. For the sake of simplicity, we often refer to a "molecule" as a group of identical compounds in the state prior to the ionization. In fact, within a single mass spectrometry experiment one can frequently observe m/z values that correspond to various adduct species of "the same molecule", as well as differing isotopic compositions of "the same molecule" (Figure 8).

After ionization, the mass analyzer separates the ions based on their m/z ratios. There are various types of mass analyzers, such as time-of-flight (QTOF), quadrupole, and Orbitrap, each with unique strengths and weaknesses. However, they all operate on the principle of using a combination of electromagnetic fields to isolate ions. Finally, the ion detector, typically integrated into the mass analyzer, measures the m/z value and intensity of the signal. While we will not delve into the technicalities, it's worth noting certain peculiarities of the measurement that are significant for subsequent MS data analysis.

Measurment of tandem mass spectra. Tandem mass spectrometry (MS/MS or MS^2) is a powerful technique enabling the detailed analysis of molecular structures by breaking down molecules into

smaller fragments and analyzing their mass-to-charge ratios (m/z). It involves using two or more subsequent mass spectrometry experiments. The first stage of mass spectrometry (**MS^1**) is out of the scope of this work; therefore, we do not discuss it here. Nevertheless, Figure 8 shows the main concept. In the second stage (**MS/MS** or **MS^2**), the instrument selects specific ions of interest termed **precursor ions**. This is realized by defining an **isolation window** of specific width that slides across the m/z range to select ions of interest. The selected m/z values can be either arbitrary (data-independent acquisition; DIA) or pre-defined beforehand (data-dependent acquisition; DDA).

Once the precursor ion is identified, it is then subjected to fragmentation. The most prominent technique to fragment the ion is collision-induced dissociation (CID). It works by accelerating the molecule towards a gas, causing it to collide with the gas molecules and recursively break into smaller substructures. The more **collision energy** is provided, the more molecular fragments are obtained as a result. The second mass spectrometer is then used to measure the **MS/MS spectrum** (also referred to as MS^2 spectrum or fragmentation spectrum), where peaks represent m/z ratios of individual fragments. Additionally, the process of selecting and fragmenting ions can be recursively repeated up to the MS^n level for any reasonable positive integer n retaining fragments.

Because the instrument can only detect ionized fragments, only a portion of substructures are recorded. For instance, a singly-charged ion broken into two fragments will form one ion and one neutral molecule (**neutral loss**) depending on the current location of the charge within the molecule. However, since "precursor ion" is in fact a group of identical molecules and the charge site is not deterministic with respect to molecule, **MS**^{2} spectrum often contains peaks for both parts with intensities reflecting their probability distribution. In particular, fragmentation spectrum often contains **precursor peak** corresponding to the whole non-fragmented precursor ion.

It is important to note that the graph-theoretical abstraction of fragmentation as a consequent removal of bonds is often, but not always, correct. For example, during CID fragmentation, a molecule can undergo rearrangement or transfer reactions leading to graph deformations, such as the formation of new rings [36, 37].

MS/MS spectra and their properties. The detector records the entire ion signal as a function of time, resulting in a mass spectrum that exhibits a continuous signal proportional to the intensity of ions relative to their mass-to-charge (m/z) ratio. This type of spectrum is known as a **profile** mass spectrum. In contrast, the instrument often produces **centroid** spectra, which undergo pre-processing via an algorithm that extracts peak information from the profile mass spectrum. The resulting signals are reported at specific m/z values and are termed **peaks** or **signals**. Also, it is a common practice to pre-process intensities such that they are represented as fractions of the maximum spectrum intensity (corresponding to the **base peak**), and are referred to as **relative intensities**.

The accuracy of the measured m/z ratios is profoundly reliant on the instrument's quality and its constituents. The two most crucial indicators of quality are the **resolution** and **accuracy** of the instrument. Resolution denotes the ability to differentiate ions with nearly identical masses, whereas accuracy indicates how close the measured m/z value is to the actual ground-truth value. Since modern instruments generally possess high separation capabilities, the resolution is often not a problem for downstream data analysis. Nonetheless, accuracy remains a pivotal concept.

Instrument vendors typically provide the accuracy of individual instruments in **ppm** (parts per million), which means that accuracy is inversely proportional to the measured mass. Specifically, a ppm of 5 would indicate that for the ground-truth m/z m, the measured value would fall within the interval $m \pm 5 \times 10^{-6} m$.

Another critical property of an instrument is its **sensitivity**. Low-sensitivity measurements may fail to detect all anticipated molecules. Conversely, high-sensitivity measurements may lead to a significant amount of **noise** – peaks that do not correspond to any actual molecules.

Annotation of MS/MS spectra. The distribution of masses provided in an MS/MS spectrum contains rich information about the underlying molecular structure. Given an MS/MS spectrum, the aim is to "arrange" the masses into a complete molecular structure. The extent to which MS/MS information is sufficient for deducing the complete structure remains a fundamental open question. However, regardless of the completeness of the structural information, a complex, high-accuracy fragmentation spectrum usually uniquely describes the molecule. The opposite statement is true only under the assumption of a similar MS experimental setup. The same compound can be fragmented in

completely different ways depending on experimental conditions such as ionization adduct species or applied collision energy. This observation rationally motivates the repeated measurement of the same compound with different instrument parameters, thereby enriching the structural information. There are multiple challenges associated with MS/MS annotation and the experimental setup, including the following examples.

First, the width of the isolation window significantly affects the information present in a fragmentation spectrum. A wide window may isolate multiple molecules of similar masses and fragment them, resulting in a single **chimeric spectrum**, which can be misleading for further annotation. Conversely, a narrow window may miss desired isotopes of the same molecule. Ultimately, the isolation window may be triggered for a wrong m/z range, resulting in a spectrum containing nothing but noise.

Second, collision energy affects the number of fragments, their size, and their structure, which in turn affects the number of peaks and their positions. Frequently, as a result of low collision energy, an MS/MS spectrum may contain only a single meaningful peak representing an unfragmented molecule. In contrast, high energy may result in too severe fragmentation, limiting structural annotation.

Finally, CID fragmentation has inherent limitations regarding the interpretation of a molecular structure. For example, it is nearly impossible to distinguish stereoisomers with sole MS/MS data. However, orthogonal sources of information, such as liquid chromatography (LC), may separate the isomers before introducing them to the mass spectrometry stage [38, 39].

5 Datasheet

Datasheets for datasets, proposed in [40], aim to serve dataset creators and consumers. They encourage creators to reflect on the creation and maintenance processes, highlighting assumptions, risks, and implications, while providing consumers with essential information to make informed decisions and avoid misuse. Below, we provide a datasheet for the MassSpecGym benchmark.

5.1 Motivation

1. **For what purpose was the dataset created?**
   The MassSpecGym benchmark was created to accelerate the process of molecule discovery and identification from tandem mass spectrometry data. To the best of our knowledge, it is the first comprehensive benchmark of its kind.
2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
   The benchmark was created as a community effort involving multiple research groups in the field of machine learning and computational metabolomics. Please see the list of authors for the complete enumeration. The idea of this benchmark set was conceived at the Dagstuhl Seminar #24181 "Computational 352 Metabolomics: Towards Molecules, Models, and their Meaning".
3. **Who funded the creation of the dataset?**
   The project was funded by national grants listed in the Acknowledgements section. Note that this article solely reflects the opinions and conclusions of its authors and not of its funders.
4. **Any other comments?**
   No other comments.

5.2 Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
   The instances of the dataset represent tandem mass spectra (measurements of molecules obtained from biological and environmental samples), each annotated with an underlying molecule.
2. **How many instances are there in total (of each type, if appropriate)?**
   The dataset contains 231 thousand instances (tandem mass spectra) annotated with 29 thousand unique molecules in total.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

   Our MassSpecGym dataset is the largest available of its kind. It is sourced from both publicly available data and our in-house measurements. The dataset provides a diverse sample of theoretically possible tandem mass spectra and molecules, richly covering molecular masses and chemical classes.

4. **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

   Each instance primarily consists of a tandem mass spectrum and the underlying molecule. Intuitively, a tandem mass spectrum is a set of 2D points (peaks), where each point represents the abundance (intensity) of molecular fragments with specific masses (m/z values). A molecule is represented as a canonical, PubChem-standardized SMILES string. Additionally, each instance contains experimental metadata and auxiliary features derived from mass spectra and molecules. Please see the supplemental materials for the complete list.

5. **Is there a label or target associated with each instance?**

   Yes, molecules define labels to be predicted from tandem mass spectra.

6. **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

   Only the instrument type and collision energy metadata features are missing for some of the instances. In these cases, the information was not deposited in the public repositories used as a source for our dataset.

7. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

   No, each instance is considered an independent measurement.

8. **Are there recommended data splits (e.g., training, development/validation, testing)?**

   Our MassSpecGym benchmark provides a data split to standardize the training and evaluation process for different models, making it easily accessible to the broad machine learning community. The data split minimizes the similarity between training and test instances by using molecule edit distance as the measure of similarity.

9. **Are there any errors, sources of noise, or redundancies in the dataset?**

   Noise in signals is inherent to mass spectrometry data. Errors and redundancies are possible in public data sources. However, our standardization pipeline aims to minimize the chances of erroneous and redundant instances.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

    The dataset is self-contained.

11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

    No, the dataset does not contain confidential data.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

    We assume that our mass spectrometry dataset does not contain any information that may cause anxiety.

5.3 Collection Process

1. **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

   The data was directly observable based on mass spectrometry measurements.

2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

   The data was collected using mass spectrometry instrumentation, pre-processed by software provided by the instrument vendors, and post-processed using open-source software (mainly matchms). The molecular labels were primarily assigned through manual human curation. These mechanisms were validated by our cleaning and quality assessment workflows.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

   Our dataset is not a sample from a larger set.

4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

   Only the authors of this work were involved in the data collection process.

5. **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

   The spectral libraries used to compile our dataset were downloaded from the official websites on May 27th, 2024.

6. **Were any ethical review processes conducted (e.g., by an institutional review board)?**

   No ethical review was conducted.

5.4 Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

   The dataset was preprocessed and cleaned using our standardization pipeline. The pipeline aims to remove noisy and corrupted spectra and ensures reliable molecular labels and metadata. Please see the supplemental materials for details. No manual labeling was performed.

2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

   The "raw" dataset was not saved as part of the MassSpecGym benchmark but can be easily downloaded from the links specified in the supplemental information.

3. **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

   The preprocessing and cleaning were largely done using the matchms library (`https://github.com/matchms/matchms`). The data processing code is publicly available on our GitHub (`https://github.com/pluskal-lab/MassSpecGym`). The only non-publicly available processing step we performed was the comparison of our mass spectra with the spectra of the commercially available NIST 23 library. This step helped us to partially clean the dataset but NIST 23 spectra were never disclosed or used in any further analyses.

4. **Any other comments?**

   No other comments.

5.5 Uses

1. **Has the dataset been used for any tasks already?**

   The dataset has already been used for the three tasks defined for the MassSpecGym benchmark: *de novo* molecule generation, molecule retrieval, and spectrum simulation.

2. **Is there a repository that links to any or all papers or systems that use the dataset?**

   The links to papers using the MassSpecGym dataset will be available through the Papers with Code resource (`https://paperswithcode.com/`).

3. **What (other) tasks could the dataset be used for?**
   The dataset can be used for other tasks related to the annotation of tandem mass spectra and, more broadly, machine learning from spectra or molecules.
4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
   To the best of our knowledge, we do not anticipate any risks or harms resulting from our dataset.
5. **Are there tasks for which the dataset should not be used?** If so, please provide a description.
   We are not aware of any such tasks. However, we anticipate that users will primarily use our dataset for the three mass spectrum annotation tasks defined in the MassSpecGym benchmark.
6. **Any other comments?**
   No other comments.

5.6 Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
   Yes, the MassSpecGym benchmark (`https://github.com/pluskal-lab/MassSpecGym`) and the MassSpecGym dataset (`https://huggingface.co/datasets/roman-bushuiev/MassSpecGym`) are publicly available.
2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**
   The training and validation data are available through the Hugging Face Datasets service. The code for training and validation is available via GitHub. An API will be available for evaluation on the test data.
3. **When will the dataset be distributed?**
   This training and validation data are publicly available at the moment of writing.
4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
   The data and code are distributed with an MIT license.
5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
   No third parties have imposed IP-based or other restrictions on the data associated with the instances.
6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
   No export controls or other regulatory restrictions apply to the dataset or to individual instances.
7. **Any other comments?**
   No other comments.

5.7 Maintenance

1. **Who will be supporting/hosting/maintaining the dataset?**
   The dataset will be supported, hosted and mainted by the authors.
2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
   The authors can be contacted through the following email addresses: roman.bushuiev@uochb.cas.cz, anton.bushuiev@cvut.cz, josef.sivic@cvut.cz, tomas.pluskal@uochb.cas.cz

3. **Is there an erratum?**
   There is no erratum. If errors are found in the future, they will be released on the project website.
4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
   Yes, if necessary to ensure high quality, the dataset will be updated with new versions on GitHub and Hugging Face Datasets.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
   The dataset does not relate to people.
6. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
   Yes, we plan to maintain the benchmark and update it to new versions, primarily when more publicly available MS/MS data becomes accessible.
7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
   Yes, users can extend, augment, build on, or contribute to the dataset and the source code by using the standard mechanisms of issues and pull requests on GitHub and Hugging Face Datasets.
8. **Any other comments?**
   No other comments.

References

[1] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[2] Swarit Jasial, Ye Hu, Martin Vogt, and Jürgen Bajorath. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research*, 5, 2016. doi: 10.12688/f1000research.8357.2.

[3] Fleming Kretschmer, Jan Seipp, Marcus Ludwig, Gunnar W Klau, and Sebastian Boecker. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv*, pages 2023–03, 2023. doi: doi.org/10.1101/2023.03.27.534311.

[4] MoNA. Massbank of north america. `https://mona.fiehnlab.ucdavis.edu/`. URL `https://mona.fiehnlab.ucdavis.edu/`.

[5] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7): 703–714, 2010. doi: 10.1002/jms.1777.

[6] Mingxun Wang, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, and et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8):828–837, Aug 2016. ISSN 1546-1696. doi: 10.1038/nbt.3597. URL `https://doi.org/10.1038/nbt.3597`.

[7] Corinna Brungs, Robin Schmid, Steffen Heuckeroth, Aninda Mazumdar, Matúš Drexler, Pavel Šácha, Pieter C Dorrestein, Daniel Petras, Louis-Felix Nothias, Radim Nencka, et al. Efficient generation of open multi-stage fragmentation mass spectral libraries. 2024. doi: 10.26434/chemrxiv-2024-l1tqh.

[8] Florian Huber, Stefan Verhoeven, Christiaan Meijer, Hanno Spreeuw, Efraín Manuel Villanueva Castilla, Cunliang Geng, Justin J. j. van der Hooft, Simon Rogers, Adam Belloum, Faruk Diblen, and Jurriaan H. Spaaks. matchms - processing and similarity evaluation of mass spectrometry

data. *Journal of Open Source Software*, 5(52):2411, 2020. doi: 10.21105/joss.02411. URL https://doi.org/10.21105/joss.02411.

[9] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16 (4):299–302, Apr 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0344-8. URL https://doi.org/10.1038/s41592-019-0344-8.

[10] NIST Mass Spectrometry Data Center. NIST20: Updates to the nist tandem and electron ionization spectral libraries. https://www.nist.gov/srd/nist-special-database-20, 2020.

[11] Niek F de Jonge, Helge Hecht, Justin JJ van der Hooft, and Florian Huber. Reproducible ms/ms library cleaning pipeline in matchms. *ChemRxiv preprint*, 2023. doi: doi:10.26434/chemrxiv-2023-l44cm.

[12] Sebastian Böcker and Florian Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. doi: 10.1093/bioinformatics/btn270. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[13] Florian Rasche, Aleš Svatoš, Ravi Kumar Maddula, Christoph Böttcher, and Sebastian Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*, 83(4): 1243–1251, 2011. doi: 10.1021/ac101825k.

[14] Sebastian Böcker and Kai Dührkop. Fragmentation trees reloaded. *Journal of cheminformatics*, 8:1–26, 2016. doi: 10.1186/s13321-016-0116-8.

[15] Volker D. Hähnke, Sunghwan Kim, and Evan E. Bolton. Pubchem chemical structure standardization. *Journal of Cheminformatics*, 10(1):36, Aug 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0293-8. URL https://doi.org/10.1186/s13321-018-0293-8.

[16] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, Russell Greiner, and David S. Wishart. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, 8(1):61, Nov 2016. ISSN 1758-2946. doi: 10.1186/s13321-016-0174-y. URL https://doi.org/10.1186/s13321-016-0174-y.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

[18] Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Emergence of molecular structures from repository-scale self-supervised learning on tandem mass spectra. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2023-kss3r-v2.

[19] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology*, 39(4):462–471, 2021. doi: 10.1038/s41587-020-0740-8.

[20] PubChem. PubChem. https://pubchem.ncbi.nlm.nih.gov/docs/statistics/. URL https://pubchem.ncbi.nlm.nih.gov/docs/statistics.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762.

[22] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1162. URL https://doi.org/10.18653/v1/p16-1162.

[23] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.*, 1(4):45024, 2020. doi: 10.1088/2632-2153/ABA947. URL https://doi.org/10.1088/2632-2153/aba947.

[24] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3391–3401, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

[25] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *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.

[26] Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J Xavier, and Connor W Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, 2023. doi: https://doi.org/10.1038/s42256-023-00708-3.

[27] Lars Ridder, Justin J. J. van der Hooft, and Stefan Verhoeven. Automatic compound annotation from mass spectrometry data using magma. *Mass Spectrometry*, 3(Special Issue 2):S0033–S0033, 2014. doi: 10.5702/massspectrometry.S0033.

[28] Jennifer N Wei, David Belanger, Ryan P Adams, and D Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science*, 5(4):700–708, 2019. doi: 10.1021/acscentsci.9b00085.

[29] Adamo Young, Fei Wang, David Wishart, Bo Wang, Hannes Röst, and Russ Greiner. FraGNNet: A Deep Probabilistic Model for Mass Spectrum Prediction, April 2024. URL http://arxiv.org/abs/2404.02360.

[30] Greg Landrum. Rdkit: Open-source cheminformatics, 2022. URL http://www.rdkit.org.

[31] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002. ISSN 0095-2338. doi: 10.1021/ci010132r.

[32] Keyulu Xu*, Weihua Hu*, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? September 2018. URL https://openreview.net/forum?id=ryGs6iA5Km.

[33] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. doi: 10.48550/arXiv.1903.02428.

[34] Roman Bushuiev. Self-supervised machine learning for the interpretation of molecular mass spectrometry data. *Master's thesis. Czech Technical University in Prague, Faculty of Information Technology*, 2023. URL https://dspace.cvut.cz/handle/10467/108811?locale-attribute=en.

[35] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, Nov 2002. ISSN 0095-2338. doi: 10.1021/ci010132r. URL https://doi.org/10.1021/ci010132r.

[36] Shandilya Mahamuni Baira, Srinivas Ragampeta, and M.V.N. Kumar Talluri. A comprehensive study on rearrangement reactions in collision-induced dissociation mass spectrometric fragmentation of protonated diphenyl and phenyl pyridyl ethers. *Rapid Communications in Mass Spectrometry*, 33(18):1440–1448, 2019. doi: https://doi.org/10.1002/rcm.8488. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/rcm.8488.

[37] Dejan Nikolić and David C. Lankin. Low energy collision-induced dissociation of azepine pictet–spengler adducts of nω-methylserotonin. *Journal of the American Society for Mass Spectrometry*, 34(2):182–192, Feb 2023. ISSN 1044-0305. doi: 10.1021/jasms.2c00247. URL https://doi.org/10.1021/jasms.2c00247.

[38] Kristina Mamko, Liudmila Kalichkina, Oleg Kotelnikov, Anna Nikulina, Natalia Dementeva, and Dmitry Novikov. Separation of cis/trans isomers of 4,5-dihydroxyimidazolidine-2-thione and 4,5-dimethoxyimidazolidine-2-thione by aqueous normal-phase hplc mode. *Chromatographia*, 83(9):1087–1093, Sep 2020. ISSN 1612-1112. doi: 10.1007/s10337-020-03926-8. URL https://doi.org/10.1007/s10337-020-03926-8.

[39] Katarzyna Bus, Jerzy Sitkowski, Wojciech Bocian, Adam Zmysłowski, Karol Ofiara, and Arkadiusz Szterk. Separation of menaquinone-7 geometric isomers by semipreparative high-performance liquid chromatography with silver complexation and identification by nuclear magnetic resonance. *Food Chemistry*, 368:130890, 2022. ISSN 0308-8146. doi: https://doi.org/10.1016/j.foodchem.2021.130890. URL https://www.sciencedirect.com/science/article/pii/S0308814621018963.

[40] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.