DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra

Montgomery Bohde [1 2] **Mrunali Manjrekar** [1] **Runzhong Wang** [1] **Shuiwang Ji** [2] **Connor W. Coley** [1]

Abstract

Mass spectrometry plays a fundamental role in elucidating the structures of unknown molecules and subsequent scientific discoveries. One formulation of the structure elucidation task is the conditional *de novo* generation of molecular structure given a mass spectrum. Toward a more accurate and efficient scientific discovery pipeline for small molecules, we present DiffMS, a formula-restricted encoder-decoder generative network that achieves state-of-the-art performance on this task. The encoder utilizes a transformer architecture and models mass spectra domain knowledge such as peak formulae and neutral losses, and the decoder is a discrete graph diffusion model restricted by the heavy-atom composition of a known chemical formula. To develop a robust decoder that bridges latent embeddings and molecular structures, we pretrain the diffusion decoder with fingerprint-structure pairs, which are available in virtually infinite quantities, compared to structure-spectrum pairs that number in the tens of thousands. Extensive experiments on established benchmarks show that DiffMS outperforms existing models on *de novo* molecule generation. We provide several ablations to demonstrate the effectiveness of our diffusion and pretraining approaches and show consistent performance scaling with increasing pretraining dataset size. DiffMS code is publicly available at https://github.com/coleygroup/DiffMS.

1. Introduction

Mass spectrometry (MS) is a fundamental part of the analytical chemistry toolkit that can assist in the identification of unknown compounds of interest collected from experiments.

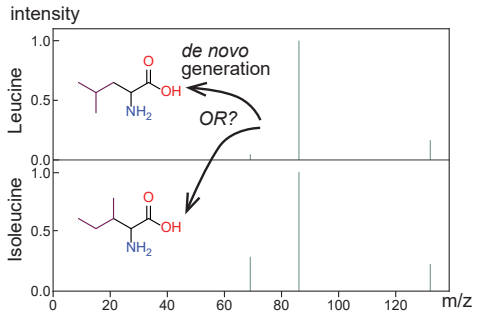

Figure 1. De novo structure generation from LC-MS/MS faces ambiguity when isobaric or isomeric compounds yield similar fragmentation spectra. In this case, the experimental spectra for leucine and isoleucine from NIST (2023) are essentially indistinguishable. It is one of many examples demonstrating that the identification of the exact structure is desirable but challenging.

Tandem mass spectrometry (MS/MS) in combination with liquid chromatography (LC) enables information-rich, high-throughput profiling of compounds, wherein complex experimental mixtures are separated in two dimensions, first by retention time (from chromatography) and then by molecule m/z (mass-to-charge ratio) in the first MS (MS1) stage. Each "precursor" molecule is then individually passed through the second MS stage (MS2) where it undergoes collision-induced dissociation and is split into a set of charged molecular fragments, each with a corresponding m/z and an intensity. Modern LC-MS/MS has enabled the discovery of many new compounds of interest, such as identifying novel bile acids in microbiome study (Quinn et al., 2020), uncovering a tire rubber-derived chemical that is toxic to coho salmon (Tian et al., 2021). There is also a growing interest in increased throughput with MS technologies, such as a high-throughput analysis of chemical reactions (Hu et al., 2024) and a systematic discovery pipeline for human metabolites (Gentry et al., 2024).

From MS1 and MS2 data alone, elucidation of the chemical structure(s) present in the original experimental sample remains challenging. Many yet-to-be-discovered metabolites have structures that do not exist in standard virtual chemical libraries (PubChem, Human Metabolome Database, etc.). The majority of observed spectra in MS-based

[1]Massachusetts Institute of Technology, Cambridge, MA, United States [2]Texas A&M University, College Station, TX, United States. Correspondence to: Connor W. Coley <ccoley@mit.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

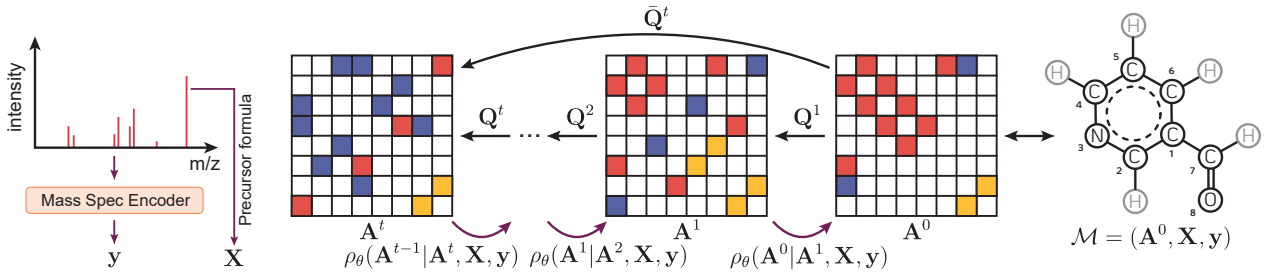

Figure 2. DiffMS tackles *de novo* molecular generation from mass spectra. We embed mass spectrum features with a transformer encoder, and assume the chemical formula is determined by off-the-shelf tools (Goldman et al., 2023c; Böcker & Dührkop, 2016) so that the numbers and types of heavy atoms (i.e. nodes in the molecular graph) is constrained. The molecular structure is represented as an adjacency matrix with one-hot encoded bond types, which in this example are single (blue), double (yellow), aromatic bonds (red) and no bond (white). The target molecular structure is generated starting from a randomly initialized adjacency matrix, which is denoised through a discrete diffusion process (Vignac et al., 2023). The trajectory used for training is created by randomly disturbing the true structure t times.

metabolomics campaigns remain unidentified and are characterized as "metabolite dark matter" (Bittremieux et al., 2022). The difficulty of the elucidation task comes from both computation and chemistry. In terms of computation, there is a large set of possible substructures to explain each measurement, creating an exponential number of structure candidates for the overall mass spectrum, i.e., an NP-hard combinatorial optimization. From the chemistry perspective, a standalone mass spectrum may be insufficient to determine a unique structure because of the ionization and fragmentation mechanisms of the instrument; as shown in Fig. 1, the spectra of two isomeric amino acids are nearly identical. While we would like to be able to determine the exact structure, from an application perspective, generating similar but not exactly matching structures is still useful to domain experts to narrow down the chemical space.

Machine learning methods have recently taken root in this space to address two key challenges in particular: 1) to learn how to fragment a given molecule, and predict the resultant mass spectrum, known as "forward" MS simulation (Murphy et al., 2023; Goldman et al., 2023a; 2024; Young et al., 2024b;a; Nowatzky et al., 2024) and 2) to take an experimental spectrum and predict the corresponding structure or a description thereof, typically as a fingerprint, SMILES, or graph representation, known as "inverse" methods (Dührkop et al., 2015; Stravs et al., 2022; Butler et al., 2023; Litsa et al., 2023; Goldman et al., 2023b).

In this paper, we focus on the "inverse" MS problem and develop a novel machine learning framework for chemical structure generation given a mass spectrum, sometimes also described as *de novo* generation. A recent study of *de novo* generation from MS shows that all of the methods tested suffer from near-zero structural accuracy (Bushuiev et al., 2024). Among prior approaches to this task are language models that are trained to convert tokenized m/z values and intensities as inputs to SMILES strings as outputs (Butler et al., 2023; Litsa et al., 2023); however, these autoregressive language models do not capture the permutation-invariant nature of mass spectra and molecules, nor do they incorporate chemical formula constraints as helpful prior knowledge. Another family of approaches utilizes intermediate representations such as scaffolds (Wang et al., 2025) or fingerprints (Stravs et al., 2022) before generating chemical structures, which are arguably more chemically interpretable and leverage additional amounts of structure-only data available, but have not necessarily led to significant performance improvement on benchmarks. Compared to the complete structural elucidation challenge of the "inverse" MS problem, the chemical formula of the unknown molecule is usually easier to determine by off-the-shelf tools from MS1 and MS2 data, utilizing tools such as SIRIUS (Böcker & Dührkop, 2016), BUDDY (Xing et al., 2023), or MIST-CF (Goldman et al., 2023c). One insight of our work is to use those available tools by taking the chemical formula as given and developing a formula-constrained (i.e., heavy atom-constrained) generation pipeline. We also find it beneficial to exploit intermediate chemical representations to enable a scaled-up pretraining stage with theoretically unlimited fingerprint-structure pairs.

To this end, we present DiffMS, a permutation-invariant diffusion model trained end-to-end for molecule generation conditioned on mass spectra (Fig. 2). DiffMS has an encoder-decoder architecture that builds upon modern transformers (Vaswani et al., 2017) and discrete graph diffusion (Vignac et al., 2023). For the encoder, we adopt the formula transformer from MIST (Goldman et al., 2023b) with pairwise modeling of neutral losses as a domain-informed inductive bias. For the decoder, we build upon the DiGress graph diffusion model (Vignac et al., 2023) using chemical formula constraints and embeddings extracted from the formula transformer as the condition to generate target

molecules. We provide empirical validation of our end-to-end model on established mass spectra *de novo* generation benchmarks (Dührkop et al., 2021b; Bushuiev et al., 2024). Additional ablation studies demonstrate the effectiveness of our pretraining-finetuning framework.

Our contributions are summarized as follows:

1. We present DiffMS, the first conditional molecular generator with formula constraints for structural elucidation from mass spectra. We demonstrate discrete diffusion as a natural methodology for conditional molecular generation that natively handles predefined heavy-atom composition and accounts for the underspecification of conditioning (i.e., the one-to-many mapping from spectrum to structure illustrated by Fig. 1).

2. We propose a pretraining-finetuning framework for training DiffMS that makes use of virtual chemical libraries with self-labeled structural conditions. Specifically, the diffusion decoder is trained on a large-scale dataset with 2.8M fingerprint-structure pairs. Our ablation studies show that downstream performance scales well with increasing fingerprint-structure pretraining dataset size, providing a promising avenue to scale the performance. We also pretrain the spectrum encoder to predict fingerprints from spectra embeddings to further boost performance of the end-to-end finetuned model.

3. On established benchmarks for *de novo* structure elucidation, DiffMS demonstrates superior performance compared to all existing baselines, achieving improved annotation accuracy and better structural similarity to the true molecule. While *de novo* generation of the exact molecular structure remains challenging, structurally close matches can offer valuable insights for domain experts (Butler et al., 2023). The broad applicability of MS underscores the potential impact of DiffMS in advancing research in chemical and biological discovery.

2. Background and Related Work

2.1. Conditional generative molecular design

Unconditional molecular generation has been well-explored in the context of AI for chemistry (Zhang et al., 2023), with methods such as Gómez-Bombarelli et al. (2018); Segler et al. (2018) leveraging autoregressive sampling with language decoders to generate SMILES representations of molecules as well as GNN architectures that generate molecular graphs atom-by-atom for either 2-dimensional (Liu et al., 2018; Li et al., 2018; Simonovsky & Komodakis, 2018) or 3-dimensional (Flam-Shepherd et al., 2022; Adams & Coley, 2022; Luo & Ji, 2022; Liu et al., 2022) graphs. Recently, Vignac et al. (2023) developed DiGress, a non-autoregressive generative model based on discrete graph diffusion. The target spaces of these generative models are generally unconditioned or loosely conditioned, for example, to generate drug-like molecules (Luo et al., 2021) or molecules with certain conformations (Roney et al., 2022).

In the context of *de novo* structural generation, however, molecular generation must be strongly conditioned on the spectral information, i.e., the fragmentation pattern itself and an inferred chemical formula. There are some efforts that try to generate structures from molecular fingerprints, which is another form of strong structural condition, including Neuraldecipher (Le et al., 2020) that learns how to decode SMILES strings from molecular fingerprints, as well as MSNovelist (Stravs et al., 2022), which proposes a fingerprint-to-SMILES long short-term memory (LSTM) network.

Both of these methods use autoregressive models that cannot strictly enforce formula constraints, while in MS, the chemical formula of the target molecule is an important inductive bias that limits the target space. In this paper, we identify discrete graph diffusion as a natural choice to incorporate formula constraints and improve Vignac et al. (2023), expanding the suite of methods in conditional molecular generation.

2.2. Inverse models for structure elucidation from spectra

Inverse models take an experimental spectrum as input and predict a relevant representation of the structure: typically, the molecular graph itself, a SMILES string, or a molecular fingerprint. DENDRAL, arguably the first expert system that applied AI to science, focuses on structural elucidation from mass spectrometry data (Lindsay et al., 1980). Recent years have seen the adoption of machine learning for a new class of inverse MS models, such as for spectrum-to-fingerprint predictions, involving either support vector machines, as in CSI:FingerID (Dührkop et al., 2015); or deep learning with transformers, as in MIST (Goldman et al., 2023b). The fingerprint, which is a binary encoding of the structure, can be further used to rank candidate structures from a chemical library. Similar elucidation goals have been pursued with other types of analytical spectra such as nuclear magnetic resonance (NMR) (Alberts et al., 2023).

However, the elucidation of structures that do not necessarily exist in any virtual chemical library requires generative techniques rather than retrieval-based techniques. MSNovelist (Stravs et al., 2022) builds an autoregressive fingerprint-to-molecule model that takes fingerprint predictions returned by CSI-FingerID and generates SMILES strings, with a decoding process that utilizes the molecular formula of the candidate compound as inferred from tools

such as SIRIUS(Böcker & Dührkop, 2016) or MIST-CF (Goldman et al., 2023c). Spec2Mol (Litsa et al., 2023) develops a SMILES autoencoder and trains a spectrum CNN encoder model, with up to four spectral channels to accept spectra collected in low or high energy and positive or negative mode, that tries to predict the corresponding SMILES embedding from the spectrum. MassGenie (Shrivastava et al., 2021) is an orthogonal effort that uses forward MS models (Allen et al., 2015; Goldman et al., 2024) to augment training datasets with *in silico* reference spectra. Toward an end-to-end pipeline for molecular generation from mass spectra, which is the most relevant to our work, Butler et al. (2023) build MS2Mol, an end-to-end language model that encodes m/z values and intensities as tokenized text input and outputs an inferred chemical formula and SMILES string in an autoregressive manner. However, their implementation is not currently available at the time of writing, preventing direct comparison. Most recently, MADGEN (Wang et al., 2025) presents a diffusion generator of chemical structures from scaffolds as a two-stage generative process, seemingly bottlenecked in terms of accuracy by scaffold prediction. In this paper, we improve upon this thread of end-to-end approaches by encoding inductive biases via spectral transformers and utilizing a pretraining-finetuning framework for an MS-conditioned diffusion generator. DiffMS has two stages like MADGEN, but is trained end-to-end during its final training step, and is heavy-atom constrained.

2.3. Diffusion generative models

Denoising diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020) has been shown to be widely effective across many tasks such as image (Song et al., 2021; Saharia et al., 2022; Karras et al., 2022) and text (Li et al., 2022; Austin et al., 2023) generation. More recently, diffusion has been applied to solve (bio)molecular generative tasks (Corso et al., 2023; Watson et al., 2023; Zeni et al., 2024).

Broadly speaking, diffusion models are generative models defined by a forward process that progressively adds noise to a sample $\boldsymbol{z}$ from data distribution $q(\boldsymbol{z}^0)$ such that $q(\boldsymbol{z}^T|\boldsymbol{z}^0)$ converges to a known prior distribution $p(\boldsymbol{z}^T)$ as $T \to \infty$. We additionally require that the noising process be Markovian such that $q(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^T|\boldsymbol{z}^0) = \prod_{t=1}^{T} q(\boldsymbol{z}^t|\boldsymbol{z}^{t-1})$. Finally, we select the forward process such that we can efficiently sample from $q(\boldsymbol{z}^t|\boldsymbol{z}^0)$.

A neural network is then trained to reverse this noising process. However, instead of predicting $p_\theta(\boldsymbol{z}^{t-1}|\boldsymbol{z}^t)$, as long as $p_\theta(\boldsymbol{z}^{t-1}|\boldsymbol{z}^t) = \int q(\boldsymbol{z}^{t-1}|\boldsymbol{z}^t, \boldsymbol{z}^0) dp_\theta(\boldsymbol{z}^0)$ is tractable, we can train the model to directly predict the denoised sample $p_\theta(\boldsymbol{z}^0|\boldsymbol{z}^t)$. To generate new samples from the model, we sample random noise from the prior distribution $p(\boldsymbol{z}^T)$, and iteratively sample from $p_\theta(\boldsymbol{z}^{t-1}|\boldsymbol{z}^t)$ until reaching $\boldsymbol{z}^0$.

Many works have also studied conditional generation with diffusion. Conditional diffusion models typically fall into two categories: classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2022). Classifier guidance uses the gradients of the log likelihood of a classifier function $p_\phi(y|\boldsymbol{z}^t)$ to guide the diffusion towards samples with class y. On the other hand, classifier-free guidance trains the denoising network directly to generate samples conditioned on class y and does not require any external classifier function. DiffMS falls under classifier-free guidance.

While diffusion models were originally designed to operate in continuous spaces, recent works have adapted denoising diffusion for discrete data modalities (Austin et al., 2023; Lou et al., 2024) and graph structured data (Vignac et al., 2023; Chen et al., 2023), both of which are relevant for molecule generation. Here, we follow the discrete diffusion settings of Austin et al. (2023) and Vignac et al. (2023).

3. Methodology

3.1. Formula-constrained molecular generation

We represent structure-spectrum pairs as $(\mathcal{M}, \mathcal{S})$, where $\mathcal{M}$ is the graph representation of the molecule with corresponding spectrum $\mathcal{S}$. The goal of *de novo* generation is to reconstruct the molecular graph $\widehat{\mathcal{M}}$ from $\mathcal{S}$. Because the molecular structure is typically underspecified given the spectrum, it is more natural to formulate *de novo* generation as predicting a ranked list of k molecules $\widehat{\boldsymbol{\mathcal{M}}}_k = (\widehat{\mathcal{M}}_1, \ldots, \widehat{\mathcal{M}}_k)$ that most closely match the given spectra.

One insight in this work is that chemical formulae represent an important physical prior that can significantly reduce the molecular search space. Formulae can be inferred from high-resolution MS1 data and isotopic traces with sufficient accuracy using tools like SIRIUS (Böcker & Dührkop, 2016) or MIST-CF (Goldman et al., 2023c), though the latter does not consider isotope distributions. To that end, we develop a formula-restricted generation using graph diffusion. In practice, we find it sufficient to model only the heavy-atoms in the graph and infer hydrogen atom placement implicitly; thus DiffMS generated molecules may differ in formula from the true molecule in hydrogen atom count.

3.2. DiffMS discrete diffusion

Let a molecular graph $\mathcal{M} = (\mathbf{A}, \mathbf{X}, \mathbf{y})$ with one-hot encoded adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n \times k}$, node features $\mathbf{X} \in \mathbb{R}^{n \times d}$ such as atom types, and graph-level structural features $\mathbf{y} \in \mathbb{R}^c$ to condition the molecule generation such as molecular fingerprint or mass spectra. Here, n is the number of heavy atoms in the molecule; $k = 5$, the number of bond types (no bond, single, double, triple, and aromatic

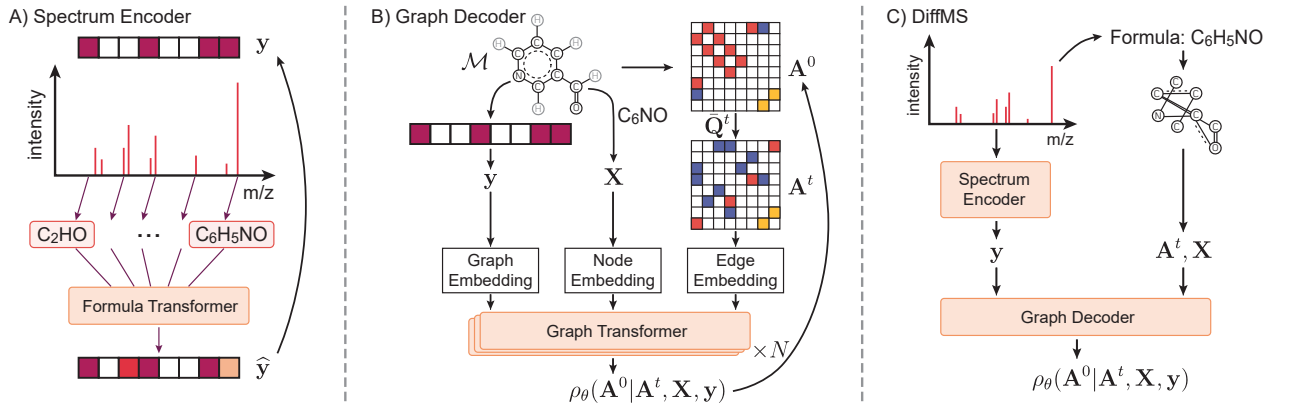

Figure 3. Model architecture of DiffMS. **A)** The spectrum encoder first assigns chemical formulae to peaks in an experimental spectrum and then learns an embedding vector through a formula transformer. The encoder is pretrained to predict Morgan fingerprints (Morgan, 1965) from spectra. **B)** The graph decoder generates the target adjacency matrix by discrete diffusion conditioned on the spectrum embedding and node (atom) features. The graph decoder is pretrained with pairs of structures and fingerprints from virtual chemical libraries. We scale up the decoder pretraining to exploit the virtually-infinite number of available fingerprint-structure pairs relative to the small number of available spectrum-structure pairs, mitigating the challenge of fingerprint-to-molecule generation found non-trivial by Le et al. (2020). **C)** DiffMS integrates the spectrum encoder and graph decoder to generate the structure annotation as a denoising process applied to a graph with randomly generated edges. It is finetuned end-to-end on labeled molecule-spectrum data.

bonds); d, the dimension of atom features; and c, the dimension of the conditional features. Because we obtain atom types from the formula, we can fix $\mathbf{X}$ and generate the adjacency matrix $\mathbf{A}$ conditioned on $\mathbf{X}$ and $\mathbf{y}$.

We define a discrete diffusion process on $\mathbf{A}$. Let $\mathbf{A}^t$ denote the value of $\mathbf{A}$ at time t. Let $\mathbf{A}^0 = \mathbf{A}$, the true molecular adjacency matrix. At each time step, $t = 1, \ldots, T$, we apply noise to each edge independently of others. Specifically, we define forward transition matrices $\left(\mathbf{Q}^1, \ldots \mathbf{Q}^T\right)$ such that $\mathbf{Q}^t_{mn} = q\left(a^t = n | a^{t-1} = m\right)$. Thus:

$$q(\mathbf{A}^t|\mathbf{A}^{t-1}) = \mathbf{A}^{t-1}\mathbf{Q}^t \tag{1}$$

Because the noise is described by a Markov transition process, we can directly sample $\mathbf{A}^t$ given $\mathbf{A}$ as:

$$q(\mathbf{A}^t|\mathbf{A}) = \mathbf{A}\bar{\mathbf{Q}}^t \tag{2}$$

Where $\bar{\mathbf{Q}}^t = \mathbf{Q}^1\mathbf{Q}^2 \ldots \mathbf{Q}^t$. Because molecular graphs are undirected, we apply noise only to the upper triangle of $\mathbf{A}$ and symmetrize the matrix. We follow the noise schedule of Vignac et al. (2023) and select

$$\bar{\mathbf{Q}}^t = \bar{\alpha}^t \boldsymbol{I} + \bar{\beta}^t \mathbf{1}_k \boldsymbol{m}^\top \tag{3}$$

where $\boldsymbol{m}$ is the marginal distribution of edge types in the training dataset and $\boldsymbol{m}^\top$ is the transpose of $\boldsymbol{m}$. This choice of $\mathbf{Q}^t$ converges to a prior distribution that is closer to the data than a uniform distribution over bond types, enabling easier training. We further select the cosine noise schedule proposed by Nichol & Dhariwal (2021):

$$\bar{\alpha}^t = \cos\left(\frac{\pi(t/T + \epsilon)}{2(1+\epsilon)}\right)^2 \tag{4}$$

with $\bar{\beta}^t = 1 - \bar{\alpha}^t$. We then define a neural network ϕ_θ that learns to predict the denoised adjacency matrix $\mathbf{A}^0$. Let $\mathcal{M}^t = (\mathbf{A}^t, \mathbf{X}, \mathbf{y})$ be the noised molecule at time t. ϕ_θ takes $\mathcal{M}^t$ as input and predicts probabilities $p_\theta(\mathbf{A}^0|\mathcal{M}^t) = \phi_\theta(\mathcal{M}^t) \in \mathbb{R}^{n \times n \times k}$. i.e., it learns to denoise $\mathbf{A}^t$ conditioned on $\mathbf{X}$, $\mathbf{y}$. We optimize this network using cross-entropy loss L between the true adjacency matrix $\mathbf{A}$ and the predicted probabilities $\hat{\mathbf{A}} = \phi_\theta(\mathcal{M}^t)$:

$$L(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{1 \le i < j \le n} \mathrm{CE}\left(a_{ij}, \hat{a}_{ij}\right) \tag{5}$$

To sample new graphs, we need to compute $p_\theta\left(\mathbf{A}^{t-1}|\mathcal{M}^t\right)$. We do so by marginalizing over the network predictions:

$$p_\theta\left(a^{t-1}_{ij}|\mathcal{M}^t\right) = \sum_{a_k} p_\theta\left(a^{t-1}_{ij}|a_{ij} = a_k, \mathcal{M}^t\right) p_\theta(a_k) \tag{6}$$

using $p_\theta\left(a^{t-1}_{ij}|a_{ij} = a_k, \mathcal{M}^t\right) = q\left(a^{t-1}_{ij}|a_{ij} = a_k, a^t_{ij}\right)$ if $q\left(a^t_{ij}|a_{ij} = a_k\right) > 0$, otherwise 0. We can then generate new graphs by sampling an initial $\mathbf{A}^T \sim \boldsymbol{m}$ and iteratively sampling from $p_\theta\left(\mathbf{A}^{t-1}|\mathcal{M}^t\right)$ until we obtain $\mathbf{A}^0$.

3.3. Model parametrization and pretraining

We use an encoder-decoder architecture to enable separate pretraining for the encoder and decoder before finetuning

the end-to-end generative model. Specifically, a spectrum encoder infers structural information from $\mathcal{S}$ and the encoder embeddings are used as the structural condition $\mathbf{y}$ for the graph diffusion decoder (Fig. 3).

For the encoder module, we use the MIST formula transformer of Goldman et al. (2023b). The encoder treats a spectrum as a set of (m/z, intensity) peaks. It embeds each peak using a predicted chemical formula assignment from SIRIUS and applies a set transformer that implicitly models pairwise neutral losses between fragments. We extract the final embedding corresponding to the precursor peak as the structural condition $\mathbf{y}$ for the diffusion decoder.

We pretrain our encoder on the same datasets used for finetuning (i.e., NPLIB1 (CANOPUS) or MassSpecGym) but now train the encoder to predict molecular fingerprints. We find that this pretraining enables the encoder to extract implicit structural information from the spectra and ensures that the encoder learns physically meaningful representations. We provide an ablation of the encoder pretraining in Sec. 4.4.

For the decoder network ϕ_θ that predicts the denoised adjacency matrix, we use a Graph Transformer (Dwivedi & Bresson, 2021). Specifically, we use separate MLPs to encode edge features $\mathbf{A}^t$, node features $\mathbf{X}$, and structural condition $\mathbf{y}$. We then apply several Graph Transformer layers before using an MLP to predict the denoised adjacency matrix $\hat{\mathbf{A}}$.

We pretrain our diffusion decoder on a dataset of fingerprint-molecule pairs. Instead of using the spectrum encoder embeddings as the structural condition $\mathbf{y}$, we directly use the molecular fingerprint to condition the molecule generation. This is closely aligned with the mass spectra *de novo* generation task, as the decoder learns to generate molecules subject to strong structural constraints. Fingerprint-molecule datasets are essentially infinite in size, providing a promising path forward to further improve model performance by increasing the pretraining dataset size. To this end, we build a pretraining dataset consisting of 2.8M fingerprint-molecule pairs sampled from DSSTox (CCTE, 2019), HMDB (Wishart et al., 2021), COCONUT (Sorokina et al., 2021), and MOSES (Polykovskiy et al., 2020) datasets. Critically, we remove *all* NPLIB1 and MassSpecGym test and validation molecules from our decoder pretraining dataset so that our evaluation on the end-to-end generation task represents a setting where the model is generating truly novel structures. Bushuiev et al. (2024) provide their own dataset of 4M molecules, but use a different exclusion criteria to prevent data leakage. We provide an ablation and analysis of performance scaling with respect to pretraining dataset size in Sec. 4.4.

4. Experiments

4.1. Evaluation metrics

We adopt the *de novo* generation metrics from Bushuiev et al. (2024):

- Top-k accuracy: measures whether the true molecule is in the top-k model predictions.
- Top-k maximum Tanimoto similarity: the structural similarity of the closest molecule to the true molecule in the top-k predictions
- Top-k minimum MCES (maximum common edge subgraph): the graph edit distance of the closest molecule to the true molecule in the top-k predictions using the distance metric proposed by Kretschmer et al. (2023).

We report metrics for $k = 1, 10$ with additional results in Appendix C. To obtain a ranked list of DiffMS predictions, we sample 100 molecules for each spectrum, remove invalid or disconnected molecules, and identify the top-k molecules based on frequency. This post-processing is also applied to baseline methods for fairest comparison.

4.2. Datasets and baselines

We evaluate DiffMS on two common open-source *de novo* generation benchmark datasets, NPLIB1 (Dührkop et al., 2021a) and MassSpecGym (Bushuiev et al., 2024). The NPLIB1 dataset is the subset of GNPS data used to train the CANOPUS tool; this term is used to disambiguate the data from the method. In order to have a fair evaluation of all methods considered, we re-implement several baseline methods to be trained only on these datasets, with modifications to the codebase if they did not have a working open-source implementation. While some papers have historically also benchmarked on the NIST20 or NIST23 datasets (NIST, 2023), this dataset is not publicly available without purchase of a license.

MSNovelist (Stravs et al., 2022) builds a fingerprint-to-SMILES LSTM decoder to predict SMILES strings from the SIRIUS-generated CSI-FingerID fingerprint. The original implementation of MSNovelist is not readily retrainable, and furthermore relies on the closed-source CSI-FingerID fingerprint. The recently developed MIST model offers an open-source replacement with reported comparative performance to CSI-FingerID (Goldman et al., 2023b). Accordingly, we re-implement a baseline model that retains the main contributions of MSNovelist, adopting the code from Zhao et al. (2024), with a 4096-bit Morgan fingerprint spectral encoder using MIST alongside a formula-guided fingerprint-to-SMILES LSTM. This fingerprint-to-SMILES decoder is trained on the same 2.8M dataset used to pretrain our diffusion decoder; therefore, unlike the original MSNovelist implementation, both the spectral encoder and LSTM decoder *never* see any test structures. We use the ranking

Table 1. De novo structural elucidation performance on NPLIB1 (Dührkop et al., 2021b) and MassSpecGym (Bushuiev et al., 2024) datasets. The best performing model for each metric is **bold** and the second best is underlined. ‡ indicates results reproduced from MassSpecGym. ∗ indicates our implementations of baseline approaches. Methods are approximately ordered by performance.

	Top-1			Top-10		
Model	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
NPLIB1						
Spec2Mol*	0.00%	27.82	0.12	0.00%	23.13	0.16
MADGEN	2.10%	20.56	0.22	2.39%	12.64	0.27
MIST + Neuraldecipher*	2.32%	12.11	**0.35**	6.11%	9.91	0.43
MIST + MSNovelist*	5.40%	14.52	0.34	11.04%	10.23	0.44
DiffMS	**8.34%**	**11.95**	**0.35**	**15.44%**	**9.23**	**0.47**
MassSpecGym						
SMILES Transformer‡	0.00%	79.39	0.03	0.00%	52.13	0.10
MIST + MSNovelist*	0.00%	45.55	0.06	0.00%	30.13	0.15
SELFIES Transformer‡	0.00%	38.88	0.08	0.00%	26.87	0.13
Spec2Mol*	0.00%	37.76	0.12	0.00%	29.40	0.16
MIST + Neuraldecipher*	0.00%	33.19	0.14	0.00%	31.89	0.16
Random Generation‡	0.00%	21.11	0.08	0.00%	18.26	0.11
MADGEN	1.31%	27.47	0.20	1.54%	16.84	0.26
DiffMS	**2.30%**	**18.45**	**0.28**	**4.25%**	**14.73**	**0.39**

methodology from MSNovelist, wherein beam search (with a width of 100) and subsequently computed log-likelihoods are used for ranking. Similarly, Spec2Mol (Litsa et al., 2023) was also retrained on the NPLIB1 and MassSpecGym datasets for fair evaluation, with only one spectral channel instead of four used for training, to alleviate restrictions on collision energy or adduct. The same ranking for candidate molecules as used for DiffMS is applied.

We also introduce a new baseline method, MIST + Neuraldecipher, that replaces the diffusion decoder in DiffMS with Neuraldecipher. Neuraldecipher encodes a molecule into a CDDD representation (Winter et al., 2019), and uses a pretrained LSTM decoder to reconstruct the SMILES string. Similar to DiffMS, we pretrain the MIST encoder on spectrum-to-fingerprint predictions, and we pretrain Neuraldecipher on fingerprint-to-molecule generation. Since MIST + Neuraldecipher uses the same pretraining-finetuning approach as DiffMS, this new baseline additionally serves as an empirical justification for our graph diffusion decoder over an LSTM-based approach.

Finally, we include a comparison to MADGEN (Wang et al., 2025). The $\text{MADGEN}_{\text{Oracle}}$ entry in Wang et al. (2025) feeds in the ground-truth scaffold which does not fall within the setting of complete *de novo* generation, and is thus not included in our evaluation. Because MADGEN uses RDKFingerprints for evaluation, as opposed to the traditional Morgan fingerprint, we exclude their Tanimoto similarities.

4.3. Results

As seen in Table 1, DiffMS outperforms baseline methods on both datasets, including more than doubling the accuracy on MassSpecGym compared to the next best method, MADGEN. While there are several baseline methods that achieve non-zero prediction accuracy on NPLIB1, only MADGEN and DiffMS generate any correct structures on MassSpecGym. NPLIB1 is inherently a less challenging dataset than MassSpecGym; given the lack of a scaffold-based split, the CANOPUS test set contains many molecules that are nearly identical (Tanimoto similarity > 0.85) to molecules in the train set (Bushuiev et al., 2024). This also explains the competitive performance of MIST + Neuraldecipher and MIST + MSNovelist, which both benefit from their ability to pretrain on these highly similar structures and learn to generate realistic structures as SMILES strings; the MSNovelist generation even more so given its formula-aware decoder. In contrast, MassSpecGym ensures that no molecules in the test set have an MCES < 10 compared to any training molecule. As such, MassSpecGym evaluation represents a more challenging and more realistic out of distribution *de novo* generation setting, which illustrates the robust performance of DiffMS across all evaluation metrics. Examples of DiffMS-generated sampled are shown in Figure 4 and Appendix E. In Appendix B, we show that even in cases where DiffMS does not recover the correct structure, it is consistently able to generate "close match" structures that are still useful to domain experts.

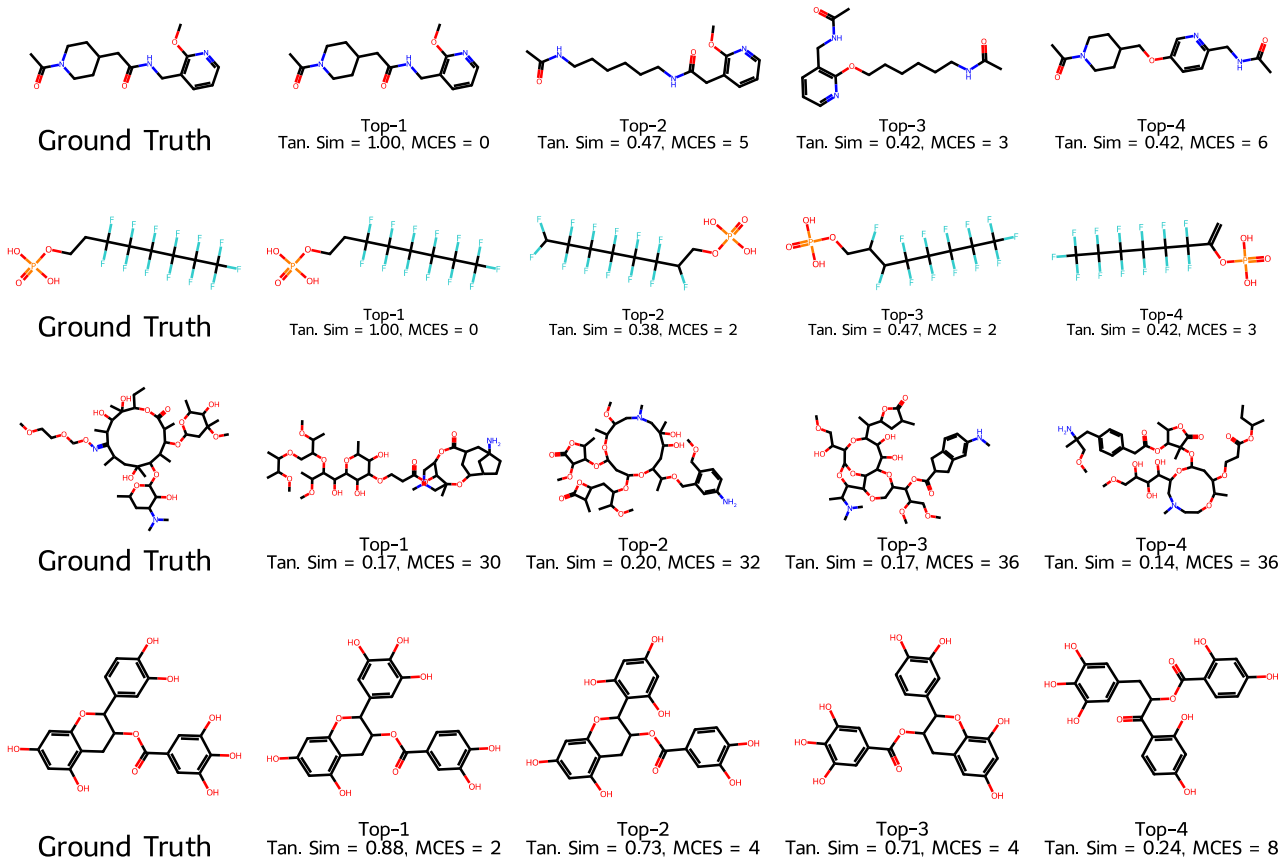

Figure 4. Ground truth molecules (left column) and DiffMS predictions (right columns) on test samples from the MassSpecGym dataset (Bushuiev et al., 2024). Tanimoto similarity and MCES metrics listed for each top-k prediction. From top to bottom, the spectra IDs are MassSpecGymID0205184, MassSpecGymID0052933, MassSpecGymID0382596, and MassSpecGymID0152454. The top two rows show cases where DiffMS successfully reconstructs the true molecule in the top-1 prediction. In the bottom two rows, DiffMS does not reconstruct the correct molecule. Additional examples can be found in Appendix E.

4.4. Pretraining Ablations

To highlight the performance gains from pretraining the DiffMS encoder and decoder, we provide several ablations. Note that comparisons in Table 1 to MIST + Neuraldecipher and MIST + MSNovelist already serve as empirical justification for DiffMS' discrete graph decoder.

Encoder Pretraining Ablation. We train DiffMS without pretraining the MIST encoder on the spectra-to-fingerprint task. As demonstrated in Table 2, encoder pretraining provides significant performance gains on the NPLIB1 dataset, nearly doubling the top-1 accuracy. Additionally, we see that even without pretraining the encoder, DiffMS generates realistic, plausible structures as indicated by the MCES and Tanimoto metrics. Nonetheless, the encoder pretraining improves the ability of the decoder to condition the diffusion on the spectra and obtain an exact match.

Decoder Pretraining Ablation. We train DiffMS with increasing decoder pretraining dataset size, starting from 0

Table 2. DiffMS performance on NPLIB1 with and without pretraining the MIST encoder on the spectrum-to-fingerprint task. The best performing model for each metric is **bold**.

Pretrain?	Accuracy ↑	MCES ↓	Tanimoto ↑
Top-1			
✗	4.36%	12.34	0.31
✓	**8.34%**	**11.95**	**0.35**
Top-10			
✗	11.46%	9.31	0.44
✓	**15.44%**	**9.23**	**0.47**

Table 3. DiffMS *de novo* structural elucidation performance on NPLIB1 (Dührkop et al., 2021b) and MassSpecGym (Bushuiev et al., 2024) datasets using MIST-CF annotated formulae and ground truth formulae. The best performing model for each metric is **bold**.

	Top-1			Top-10		
Formulae	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
NPLIB1						
MIST-CF Formulae	7.03%	**11.81**	**0.36**	14.98%	9.39	**0.48**
True Formulae	**8.34%**	11.95	0.35	**15.44%**	**9.23**	0.47
MassSpecGym						
MIST-CF Formulae	1.86%	**17.83**	0.27	4.10%	**13.71**	**0.40**
True Formulae	**2.30%**	18.45	**0.28**	**4.25%**	14.73	0.39

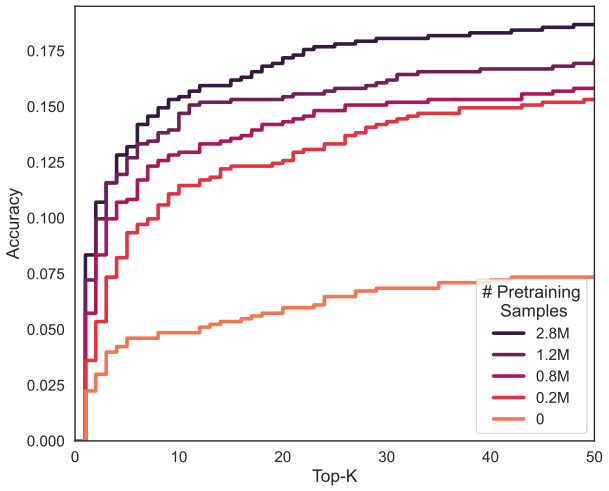

Figure 5. NPLIB1 top-k accuracy for DiffMS pretrained on increasingly large fingerprint-to-molecule datasets. Additional metrics available in Table 5 in the Appendix.

molecules (i.e., no pretraining) up to the full pretraining dataset of 2.8M molecules. As shown in Fig. 5, any amount of decoder pretraining offers a significant increase in performance. Additionally, we observe good performance scaling with increasing pretraining dataset size. Since fingerprint-to-molecule datasets are essentially infinite in size, this provides an avenue to continue scaling DiffMS' performance by building even larger and more chemically comprehensive pretraining datasets.

Additional results and figures for encoder and decoder pretraining ablations can be found in Appendix C.

4.5. Formula Inference Ablation

In many real-world elucidation settings, chemists may know the true chemical formula of the target compound *a priori* or be able to determine the true formula using auxiliary methods. However, chemical formulae can also be predicted from the spectrum with high accuracy by out-of-the-box formula annotation tools (Goldman et al., 2023c; Xing et al., 2023; Böcker & Dührkop, 2016). In this section, we broaden the structural elucidation challenge and investigate the ability of DiffMS to rely on MIST-CF (Goldman et al., 2023c) formula predictions to test its performance in settings where the true chemical formula is unknown.

For each spectrum, we predict the top 5 most likely formulae using MIST-CF. We then generate candidate structures for each of the predicted formulae. To have a fair comparison, we still generate 100 total molecules, split across the 5 predicted formulae. As shown in Table 3, DiffMS still has strong performance even when relying on formula annotation tools to supply the formula. While the elucidation accuracy is slightly lower, the MCES and Tanimoto metrics are actually better in some cases using the MIST-CF predicted formulae. Intuitively, sampling molecules with different formulae gives us higher diversity and thus a better chance of getting a "close" structure.

5. Conclusion

In this work, we propose DiffMS, a conditional molecule generative model with formula constraints for structural elucidation from mass spectra. We develop a pretraining-finetuning framework for separate spectra encoder and graph diffusion decoders that makes use of extensive fingerprint-molecule datasets and ensures the spectrum encoder learns to extract physically meaningful representations from mass spectra. We show that DiffMS achieves state-of-the-art results across common *de novo* generation benchmarks, and provide several ablations to demonstrate the effectiveness of our contributions and the potential to further improve performance by scaling pretraining.

Acknowledgments

This work was partly sponsored by DSO National Laboratories in Singapore (to C.W.C.), the MIT Summer Re-

search Program (MSRP), National Science Foundation under grant IIS-2243850 (to S.J.), and ARPA-H under grant 1AY1AX000053 (to S.J.).

Impact Statement

The advancement of computational tools for structure elucidation will aid in the identification of unknown molecules, including metabolites as biomarkers for diagnostic applications or improving understanding of biology. There are many potential beneficial societal consequences of our work and very few potential negative ones, none of which warrant elaboration here.

References

Adams, K. and Coley, C. W. Equivariant shape-conditioned generation of 3d molecules for ligand-based drug design. *arXiv preprint arXiv:2210.04893*, 2022.

Alberts, M., Zipoli, F., and Vaucher, A. C. Learning the language of nmr: Structure elucidation from nmr spectra using transformer models. In *NeurIPS 2023 Workshop*, 2023.

Allen, F., Greiner, R., and Wishart, D. Competitive fragmentation modeling of esi-ms/ms spectra for putative metabolite identification. *Metabolomics*, 11:98–110, 2015.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

Bittremieux, W., Wang, M., and Dorrestein, P. C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics*, 18 (12):94, 2022.

Böcker, S. and Dührkop, K. Fragmentation trees reloaded. *Journal of cheminformatics*, 8:1–26, 2016.

Bushuiev, R., Bushuiev, A., de Jonge, N. F., Young, A., Kretschmer, F., Samusevich, R., Heirman, J., Wang, F., Zhang, L., Dührkop, K., Ludwig, M., Haupt, N. A., Kalia, A., Brungs, C., Schmid, R., Greiner, R., Wang, B., Wishart, D. S., Liu, L.-P., Rousu, J., Bittremieux, W., Rost, H., Mak, T. D., Hassoun, S., Huber, F., van der Hooft, J. J. J., Stravs, M. A., Böcker, S., Sivic, J., and Pluskal, T. Massspecgym: A benchmark for the discovery and identification of molecules, 2024. URL https://arxiv.org/abs/2410.23326.

Butler, T., Frandsen, A., Lightheart, R., Bargh, B., Taylor, J., Bollerman, T., Kerby, T., West, K., Voronov, G., Moon, K., et al. Ms2mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv*, 2023.

CCTE, E. Distributed Structure-Searchable Toxicity (DSSTox) Database. 4 2019. doi: 10.23645/epacomptox.5588566.v7. URL https://epa.figshare.com/articles/dataset/Chemistry_Dashboard_Data_DSSTox_Identifiers_Mapped_to_CAS_Numbers_and_Names/5588566.

Chen, X., He, J., Han, X., and Liu, L.-P. Efficient and degree-guided graph generation via discrete diffusion modeling, 2023. URL https://arxiv.org/abs/2305.04111.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.

Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., Petras, D., Gerwick, W. H., Rousu, J., Dorrestein, P. C., and Böcker, S. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 39(4):462–471, Apr 2021a. ISSN 1546-1696. doi: 10.1038/s41587-020-0740-8. URL https://doi.org/10.1038/s41587-020-0740-8.

Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., Petras, D., Gerwick, W. H., Rousu, J., Dorrestein, P. C., et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology*, 39(4): 462–471, 2021b.

Dwivedi, V. P. and Bresson, X. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Flam-Shepherd, D., Zhigalin, A., and Aspuru-Guzik, A. Scalable fragment-based 3d molecular design with reinforcement learning. *arXiv preprint arXiv:2202.00658*, 2022.

Gentry, E. C., Collins, S. L., Panitchpakdi, M., Belda-Ferre, P., Stewart, A. K., Carrillo Terrazas, M., Lu, H.-h., Zuffa, S., Yan, T., Avila-Pacheco, J., et al. Reverse metabolomics for the discovery of chemical structures from humans. *Nature*, 626(7998):419–426, 2024.

Goldman, S., Bradshaw, J., Xin, J., and Coley, C. Prefix-tree decoding for predicting mass spectra from molecules. *Advances in Neural Information Processing Systems*, 36: 48548–48572, 2023a.

Goldman, S., Wohlwend, J., Stražar, M., Haroush, G., Xavier, R. J., and Coley, C. W. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, 2023b.

Goldman, S., Xin, J., Provenzano, J., and Coley, C. W. Mist-cf: Chemical formula inference from tandem mass spectra. *Journal of Chemical Information and Modeling*, 64(7):2421–2431, 2023c.

Goldman, S., Li, J., and Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks. *Analytical Chemistry*, 96(8):3419–3428, 2024.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Hu, M., Yang, L., Twarog, N., Ochoada, J., Li, Y., Vrettos, E. I., Torres-Hernandez, A. X., Martinez, J. B., Bhatia, J., Young, B. M., et al. Continuous collective analysis of chemical reactions. *Nature*, 636(8042):374–379, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

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

Le, T., Winter, R., Noé, F., and Clevert, D.-A. Neuraldecipher–reverse-engineering extended-connectivity fingerprints (ecfps) to their molecular structures. *Chemical science*, 11(38):10378–10389, 2020.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018.

Lindsay, R. K., Buchanan, B., Feigenbaum, E., and Lederberg, J. *Applications of artificial intelligence for organic chemistry: the DENDRAL project*. McGraw-Hill Companies, 1980.

Litsa, E. E., Chenthamarakshan, V., Das, P., and Kavraki, L. E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry*, 6(1):132, 2023.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020.

Liu, M., Luo, Y., Uchino, K., Maruhashi, K., and Ji, S. Generating 3d molecules for target protein binding. In *International Conference on Machine Learning*, 2022.

Liu, Q., Allamanis, M., Brockschmidt, M., and Gaunt, A. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31, 2018.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*, 2019.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Luo, S., Guan, J., Ma, J., and Peng, J. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Luo, Y. and Ji, S. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International conference on learning representations (ICLR)*, 2022.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Murphy, M., Jegelka, S., Fraenkel, E., Kind, T., Healey, D., and Butler, T. Efficiently predicting high resolution mass spectra with graph neural networks. In *International Conference on Machine Learning*, pp. 25549–25562. PMLR, 2023.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

NIST. NIST standard reference database. National Institute of Standards and Technology, 2023. URL https://www.nist.gov/srd.

Nowatzky, Y., Russo, F., Lisec, J., Kister, A., Reinert, K., Muth, T., and Benner, P. Fiora: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *bioRxiv*, pp. 2024–04, 2024.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., and Zhavoronkov, A. Molecular sets (moses): A benchmarking platform for molecular generation models, 2020. URL https://arxiv.org/abs/1811.12823.

Quinn, R. A., Melnik, A. V., Vrbanac, A., Fu, T., Patras, K. A., Christy, M. P., Bodai, Z., Belda-Ferre, P., Tripathi, A., Chung, L. K., et al. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature*, 579(7797):123–129, 2020.

Roney, J. P., Maragakis, P., Skopp, P., and Shaw, D. E. Generating realistic 3d molecules with an equivariant conditional likelihood model. 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Segler, M. H., Kogej, T., Tyrchan, C., and Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4 (1):120–131, 2018.

Shrivastava, A. D., Swainston, N., Samanta, S., Roberts, I., Wright Muelas, M., and Kell, D. B. Massgenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793, 2021.

Simonovsky, M. and Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, pp. 412–422. Springer, 2018.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., and Steinbeck, C. Coconut online: Collection of open natural products database. *Journal of Cheminformatics*, 13(1), Jan 2021. doi: 10.1186/s13321-020-00478-9.

Stravs, M. A., Dührkop, K., Böcker, S., and Zamboni, N. Msnovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022.

Tian, Z., Zhao, H., Peter, K. T., Gonzalez, M., Wetzel, J., Wu, C., Hu, X., Prat, J., Mudrock, E., Hettinger, R., et al. A ubiquitous tire rubber–derived chemical induces acute mortality in coho salmon. *Science*, 371(6525):185–189, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Neural Info. Process. Systems*, volume 30, 2017.

Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., and Frossard, P. Digress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Wang, Y., Chen, X., Liu, L., and Hassoun, S. MADGEN: Mass-spec attends to de novo molecular generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=78tc3EiUrN.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V.,

Lauko, A., Bortoli, V. D., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089 – 1100, 2023. URL https://api.semanticscholar.org/CorpusID:271161349.

Winter, R., Montanari, F., Noé, F., and Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.*, 10:1692–1701, 2019. doi: 10.1039/C8SC04175J. URL http://dx.doi.org/10.1039/C8SC04175J.

Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V., Varshavi, D., Varshavi, D., Allen, D., Arndt, D., Khetarpal, N., Sivakumaran, A., Harford, K., Sanford, S., Yee, K., Cao, X., Budinski, Z., Liigand, J., Zhang, L., Zheng, J., Mandal, R., Karu, N., Dambrova, M., Schiöth, H., Greiner, R., and Gautam, V. Hmdb 5.0: the human metabolome database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1062. URL https://doi.org/10.1093/nar/gkab1062.

Xing, S., Shen, S., Xu, B., Li, X., and Huan, T. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nat. Methods*, 20(6):881–890, June 2023.

Young, A., Röst, H., and Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416, 2024a.

Young, A., Wang, F., Wishart, D., Wang, B., Röst, H., and Greiner, R. Fragnnet: A deep probabilistic model for mass spectrum prediction. *arXiv preprint arXiv:2404.02360*, 2024b.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Shysheya, S., Crabbé, J., Sun, L., Smith, J., Nguyen, B., Schulz, H., Lewis, S., Huang, C.-W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Tomioka, R., and Xie, T. Mattergen: a generative model for inorganic materials design, 2024. URL https://arxiv.org/abs/2312.03687.

Zhang, X., Wang, L., Helwig, J., Luo, Y., Fu, C., Xie, Y., ..., and Ji, S. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

Zhao, K., Liu, Y., Dian, L., Sun, S., and Cui, X. How to train your neural network for molecular structure generation from mass spectra? In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 817–822. IEEE, December 2024.

Table 4. Additional evaluation of molecule validity, and percentage above domain-expert-defined Tanimoto thresholds on NPLIB1 (Dührkop et al., 2021b) and MassSpecGym (Bushuiev et al., 2024) *de novo* generation datasets. The best performing model for each metric is **bold** and the second best is underlined. Definitions of meaningful match (Tanimoto similarity ≥ 0.4) and close match (Tanimoto similarity ≥ 0.675) are taken from Butler et al. (2023).

	Overall	Top-1		Top-10	
Model	% Valid ↑	% Meaningful match ↑	% Close match ↑	% Meaningful match ↑	% Close match ↑
			NPLIB1		
Spec2Mol	66.5%	0.00%	0.00%	0.00%	0.00%
MIST + Neuraldecipher	91.11%	29.30%	7.33%	41.39%	12.82%
MIST + MSNovelist	98.60%	**32.90%**	11.78%	44.79%	19.02%
DiffMS	**100.0%**	27.40%	**12.83%**	**46.45%**	**22.04%**
			MassSpecGym		
Spec2Mol	68.5%	0.0%	0.0%	0.0%	0.0%
MIST + Neuraldecipher	81.78%	0.29%	0.01%	0.39%	0.09%
MIST + MSNovelist	98.58%	0.66%	0.00%	1.92%	0.00%
DiffMS	**100.0%**	**12.41%**	**3.78%**	**32.47%**	**6.73%**

A. Experimental Details

For node features $\mathbf{X}$, we use a one-hot encoding of atom types, $\mathbf{X} \in \mathbb{R}^{n \times d}$, where d is the number of different atom types in the dataset.

For pretraining the decoder, we use 2048-bit Morgan fingerprints with radius 2 for the structural conditioning $\mathbf{y} \in \mathbb{R}^{2048}$. We use the same training objective as the end-to-end finetuning, i.e., minimizing the cross-entropy loss between the denoised adjacency matrix $\hat{\mathbf{A}}$ and the true adjacency matrix, $\hat{\mathbf{A}}$. We build a decoder pretraining datset consisting of 2.8M fingerprint-molecule pairs sampled from DSSTox (CCTE, 2019), HMDB (Wishart et al., 2021), COCONUT (Sorokina et al., 2021), and MOSES (Polykovskiy et al., 2020) datasets. We pretrain the decoder for 100 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017) and a cosine annealing learning rate scheduler (Loshchilov & Hutter, 2019).

We pretrain the encoder on the same dataset used for finetuning (i.e. NPLIB1, MassSpecGym), which are orders of magnitude smaller than the decoder pretraining dataset. For encoder pretraining, we use the multi-objective loss settings of Goldman et al. (2023b). We pretrain the encoder for 100 epochs using the RAdam optimizer (Liu et al., 2020).

We finetune the end-to-end model using cross-entropy loss and no auxiliary training objectives, i.e. only the denoising diffusion objective. We use the AdamW optimizer with cosine annealing learning rate schedule for finetuning. We finetune DiffMS for 50 epochs on NPLIB1 and 15 epochs on MassSpecGym.

DiffMS is a relatively lightweight model, and all experiments were run on NVIDIA 2080ti GPUs with 12 GB of memory. On these GPUs, finetuning DiffMS takes 1.45 minutes per epoch on CANOPUS and 46 minutes per epoch on MassSpecGym. It takes 4 minutes on average to generate 100 samples from DiffMS.

B. Additional Results

As an addendum to the evaluations in Table 1, we provide some additional metrics to further contextualize DiffMS performance. Firstly, we evaluate the percentage of model samples that correspond to valid molecules. Additionally, we adopt the domain-expert thresholds put forth by MS2Mol (Butler et al., 2023), where we evaluate whether candidate molecules were a "meaningful" match in structural similarity, having a Tanimoto similarity of 0.4 or greater; or a "close match" in structural similarity, having a Tanimoto similarity of 0.675 or greater. We omit MADGEN and the baseline methods from Bushuiev et al. (2024) as they do not report these metrics.

As shown in Table 4, 100% of DiffMS samples are valid molecules. This is directly enforced because of our graph-based representation. In contrast, SMILES strings generated by baseline methods may not correspond to a valid structure. We find that DiffMS consistently achieves higher meaningful and close match rates than baseline methods. Impressively, DiffMS achieves over 32 times more meaningful matches in the top-10 predictions than the next best baseline on MassSpecGym. These results show that while generating exact matches continues to be a challenging task for *de novo* structural elucidation, DiffMS is able to generate meaningful structural matches at a high rate.

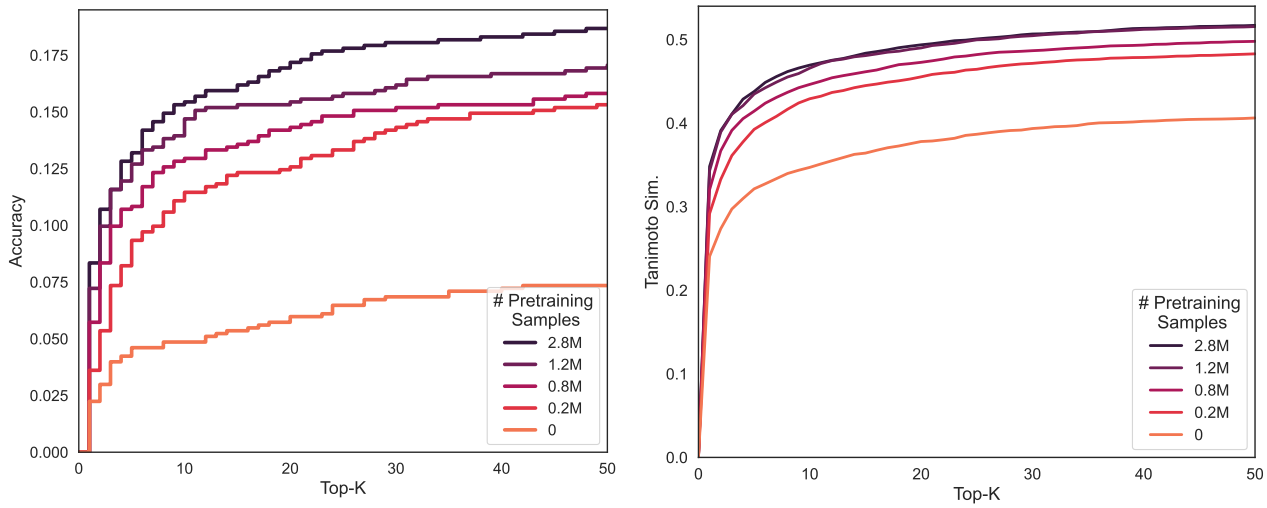

Figure 6. Annotation accuracy (left) and Tanimoto similarity (right) on the NPLIB1 dataset for DiffMS pretrained on increasingly large pretraining datasets.

Table 5. DiffMS performance on NPLIB1 for DiffMS pretrained on increasingly large fingerprint-to-molecule datasets. The best performing model for each metric is **bold** and second best is underlined.

# Pretraining Structures	Top-1			Top-10		
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
0	2.22%	15.37	0.22	4.86%	12.06	0.34
0.2M	3.61%	13.22	0.28	10.71%	9.85	0.41
0.8M	5.60%	13.02	0.30	12.70%	9.86	0.44
1.2M	7.22%	**11.63**	0.33	14.69%	**9.23**	0.43
2.8M	**8.34%**	11.95	**0.35**	**15.44%**	**9.23**	**0.47**

C. Ablations

C.1. Additional Pretraining Ablation Results

In this section, we provide additional results for the ablation studies in Sec. 4.4. Table 5 and Fig. 7 demonstrate DiffMS' performance scaling with respect to increasingly large decoder pretraining datasets, and Fig 6 shows the impact of pretraining the spectra encoder.

C.2. Prior Distribution Ablations

In this section we provide an additional ablation study to justify the choice of the marginal prior distribution. Specifically, we compare with two alternative prior distributions: the "empty" distribution, consisting of no bonds, and the "fully connected" distribution, consisting of all single bonds. As shown in Table 6, the marginal distribution performs best, though the empty distribution is not far behind. Intuitively, the empty distribution is close to the marginal distribution as molecular graphs are typically very sparse. These results support our intuitions that having a prior distribution that is closer to the data distribution results in better performance.

D. Formulae Annotation Study

In this section, we provide additional experiments using MIST-CF (Goldman et al., 2023b) and BUDDY (Xing et al., 2023) for formula annotation on the NPLIB1 and MassSpecGym datasets. Specifically, we use BUDDY and MIST-CF to predict the top-5 most likely formulae for each spectra in the test sets and measure the accuracy of these formula annotations.

As shown in Table 7, MIST-CF and BUDDY both achieve good performance on NPLIB1, where MIST-CF achieves over

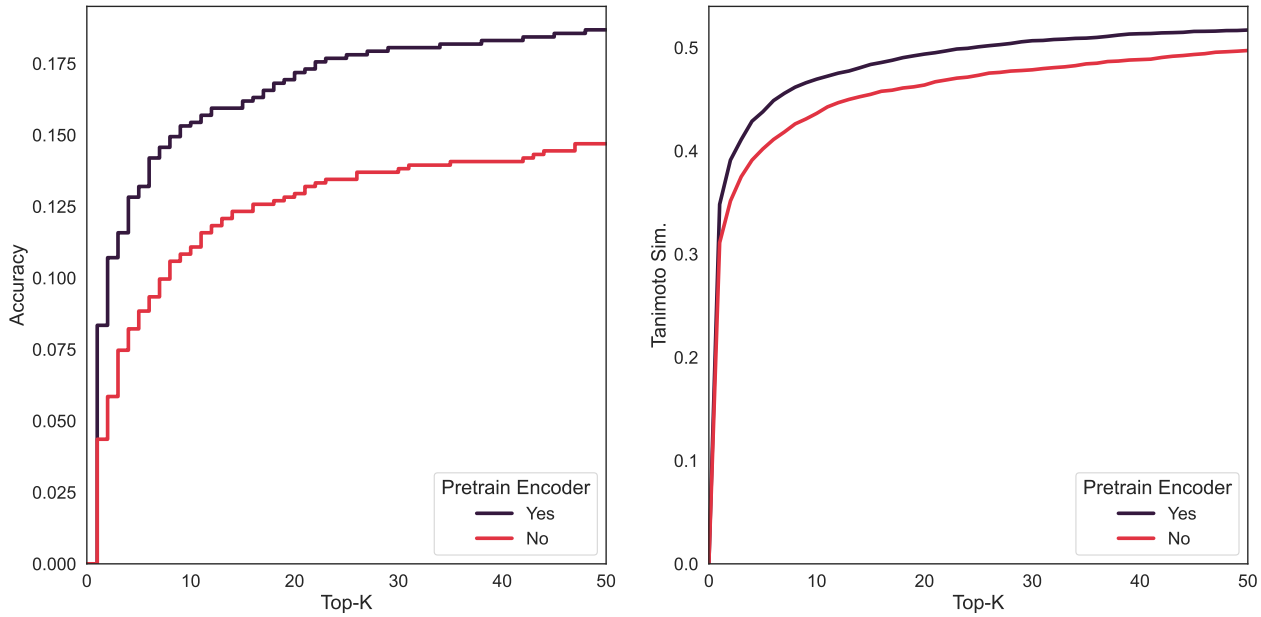

Figure 7. Annotation accuracy (left) and Tanimoto similarity (right) on the NPLIB1 dataset for DiffMS with and without encoder pretraining.

Table 6. DiffMS performance on NPLIB1 with different prior distributions. The best performing model for each metric is **bold** and second best is underlined.

Prior Distribution	Top-1			Top-10		
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
Fully Connected	3.36%	12.67	0.28	7.60%	9.56	0.4
Empty	6.60%	**11.55**	0.34	14.94%	**9.07**	**0.47**
Marginal	**8.34%**	11.95	**0.35**	**15.44%**	9.23	**0.47**

90% top-5 accuracy. However, both methods struggle on MassSpecGym, underscoring the difficulty of this dataset. It is important to note that neither NPLIB1 nor MassSpecGym include MS1 data, such as precursor m/z, which can aid in deriving accurate formula annotations. As such, these formula annotation accuracies are likely lower than what could be achieved in end-to-end elucidation workflows.

Table 7. Formula annotation accuracy for MIST-CF (Goldman et al., 2023c) and BUDDY(Xing et al., 2023) on the NPLIB1 (Böcker & Dührkop, 2016) and MassSpecGym (Bushuiev et al., 2024) datasets. The best performing method for each metric is **bold**.

Model	NPLIB1		MassSpecGym	
	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
BUDDY	78%	83%	**59%**	**71%**
MIST-CF	**84%**	**92%**	48%	69%

E. DiffMS Generated Molecules

E.1. NPLIB1 Molecules

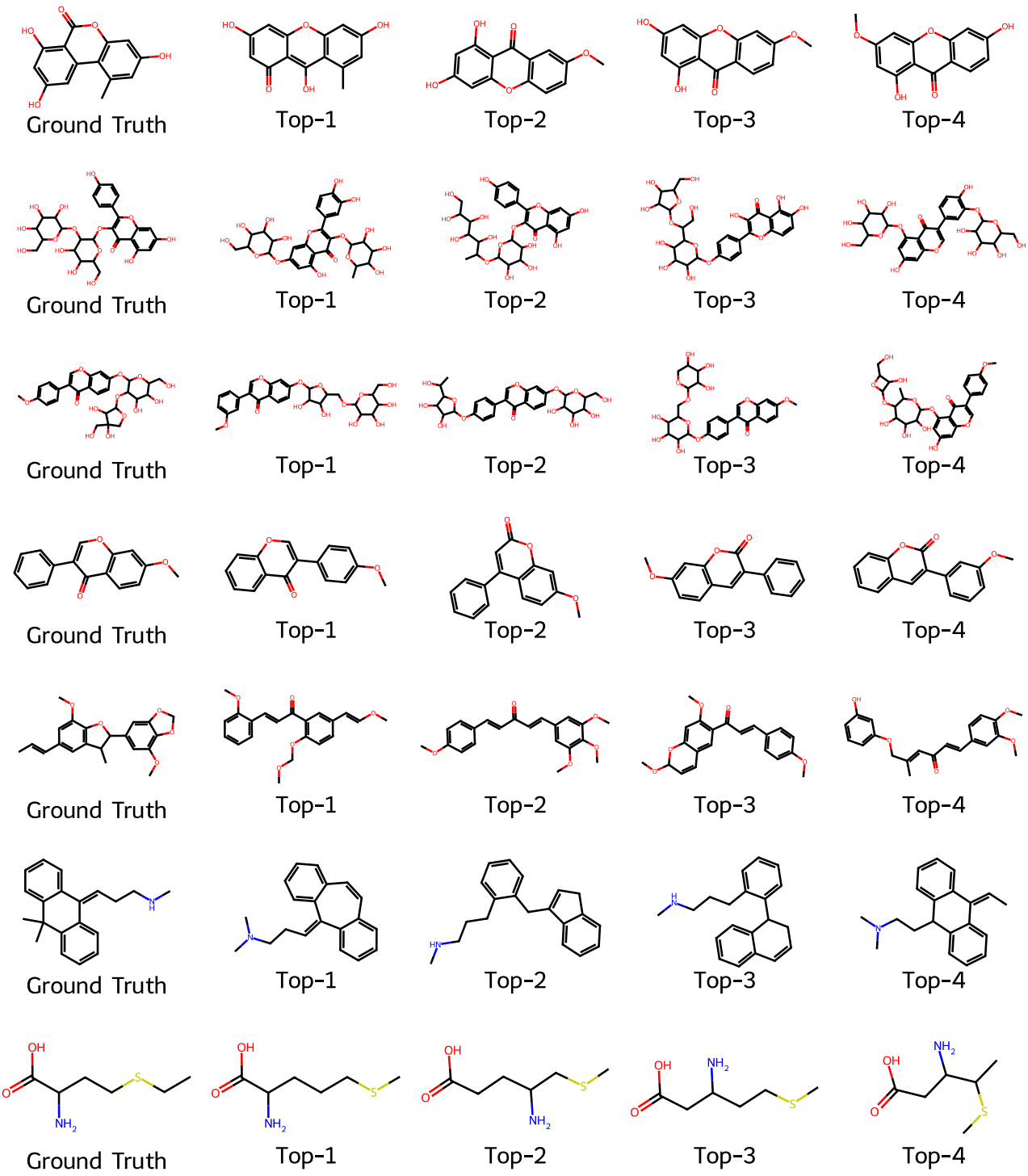

Figure 9. Negative (failure) test samples from the NPLIB1 dataset (Dührkop et al., 2021b). Ground truth molecules (left column) and DiffMS predictions (right columns).

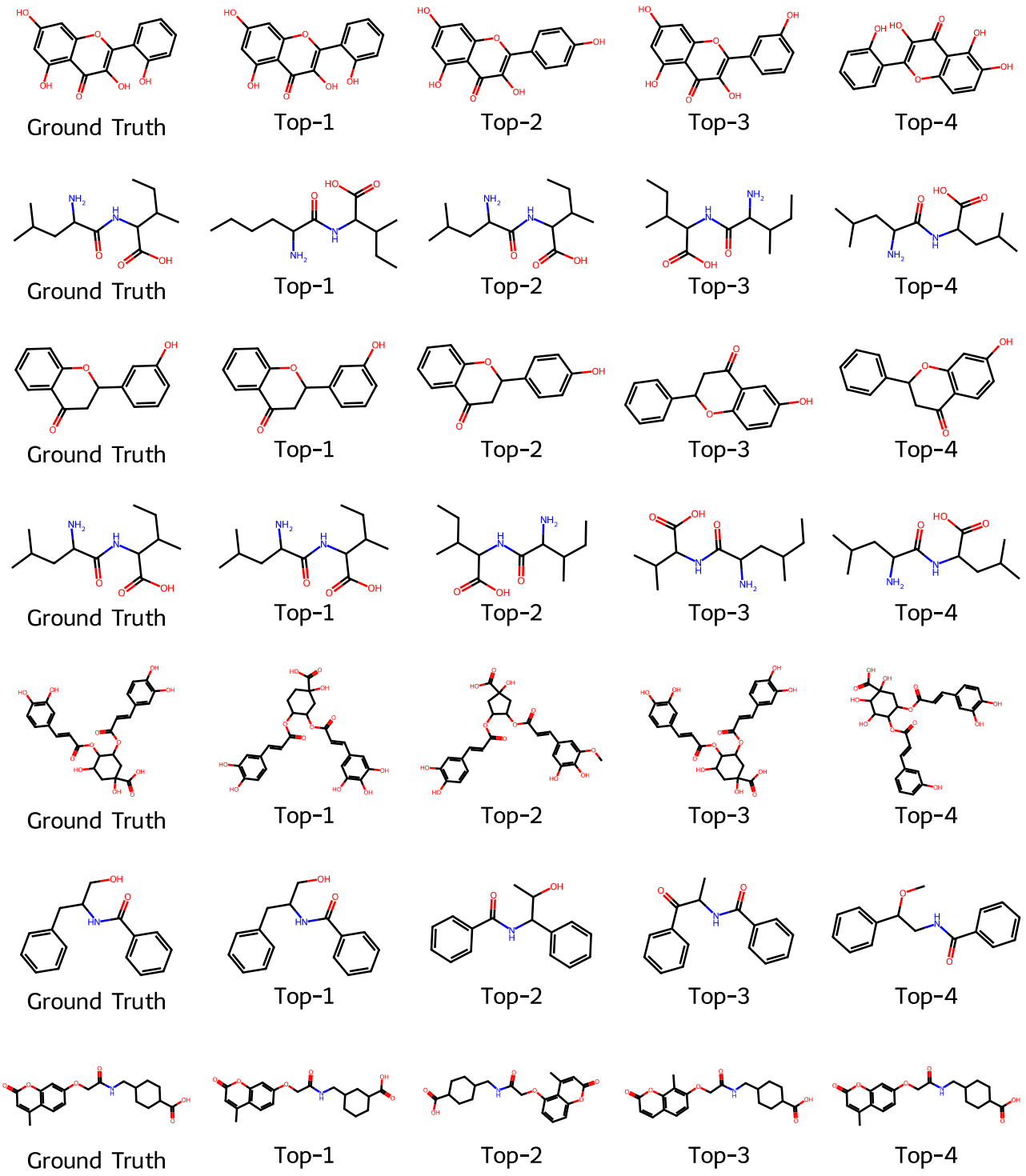

Figure 8. Positive (correct) test samples from the NPLIB1 dataset (Dührkop et al., 2021b). Ground truth molecules (left column) and DiffMS predictions (right columns).

E.2. MassSpecGym Molecules

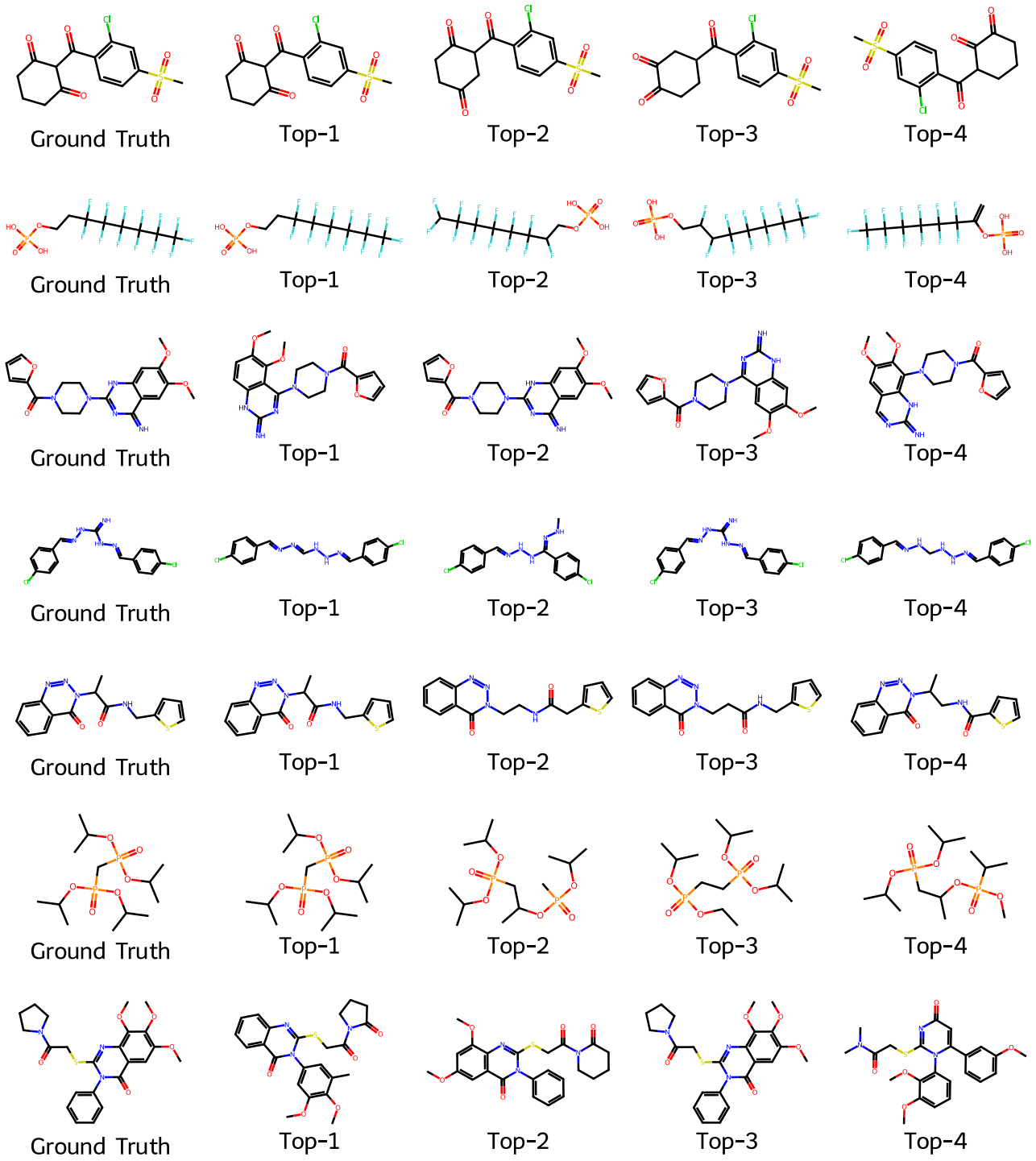

Figure 10. Positive (correct) test samples from the MassSpecGym dataset (Bushuiev et al., 2024). Ground truth molecules (left column) and DiffMS predictions (right columns).

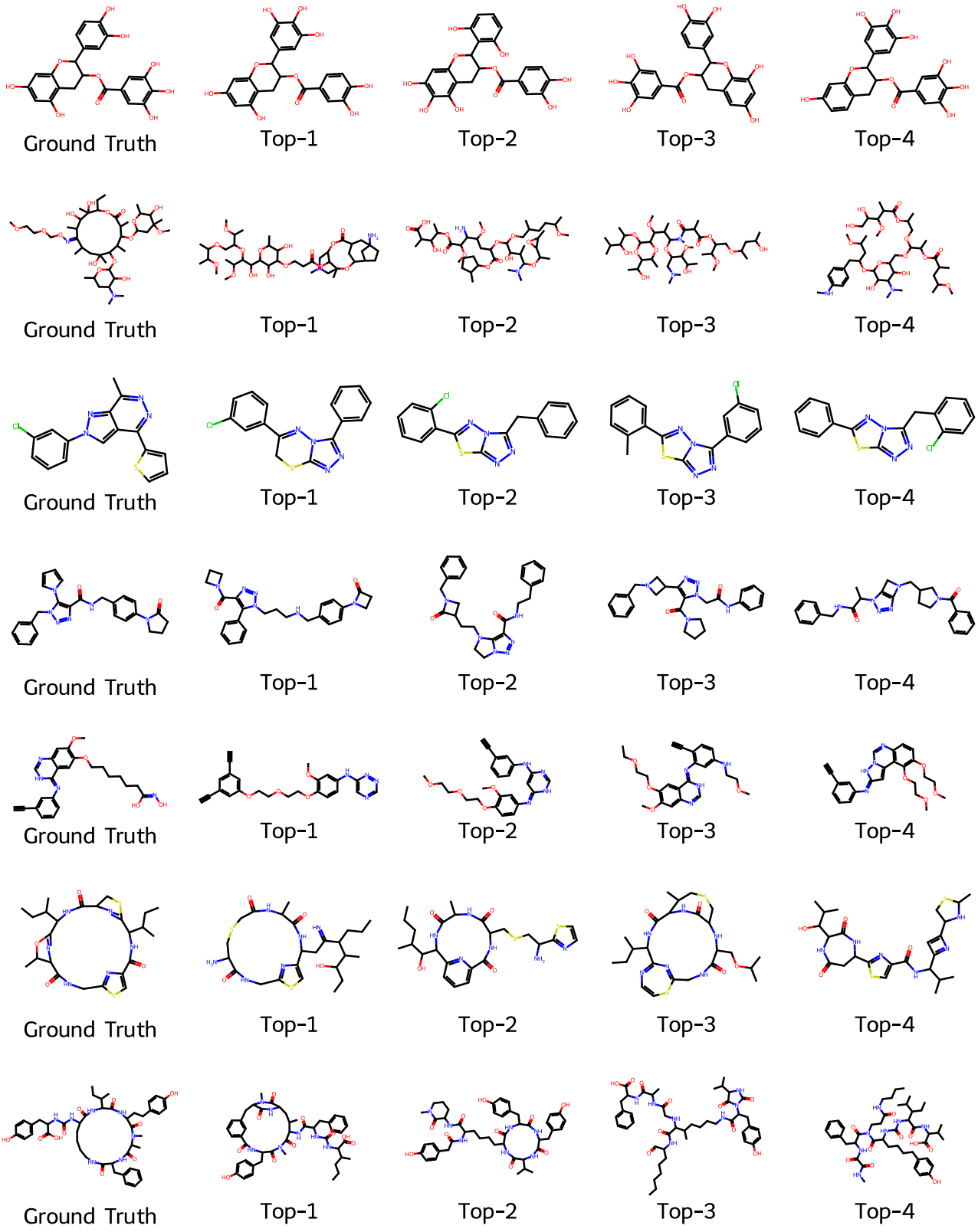

Figure 11. Negative (failure) test samples from the MassSpecGym dataset (Bushuiev et al., 2024). Ground truth molecules (left column) and DiffMS predictions (right columns).