One Small Step with Fingerprints, One Giant Leap for *De Novo* Molecule Generation from Mass Spectra

Neng Kai Nigel Neo
nnengkai@dso.org.sg

Lim Jing

Ngoui Yong Zhau Preston

Koh Xue Ting Serene

Shen Bingquan
SBingqua@dso.org.sg

DSO National Laboratories
Singapore

Abstract

A common approach to the *de novo* molecular generation problem from mass spectra involves a two-stage pipeline: (1) encoding mass spectra into molecular fingerprints, followed by (2) decoding these fingerprints into molecular structures. In our work, we adopt MIST [5] as the encoder and MOLFORGE [16] as the decoder, leveraging additional training data to enhance performance. We also threshold the probabilities of each fingerprint bit to focus on the presence of substructures. This results in a tenfold improvement over previous state-of-the-art methods, generating top-1 31% / top-10 40% of molecular structures correctly from mass spectra in MassSpecGym [2]. We position this as a strong baseline for future research in *de novo* molecule elucidation from mass spectra.

1 Introduction

Mass spectrometry (MS) is a foundational technique in analytical chemistry, widely employed for the structural characterisation of small molecules. MS analysis has been used in monitoring reaction intermediates in catalytic cycles and characterising degradation products that govern material longevity. However, the interpretation of mass spectra to elucidate the structure of unknown compounds remains a significant bottleneck, requiring extensive domain knowledge and time-consuming manual effort. Automated solutions to this problem, known as *de novo* molecule generation from mass spectra, have been proposed over the years. One general approach uses a two-stage pipeline of (1) encoding mass spectra into molecular fingerprints (FP), followed by (2) decoding these fingerprints into molecular structures [1]. Hence, choosing appropriate encoders and decoders is important.

Previous approaches utilising this pipeline suffer from poor performance due to the dependency of the fingerprint decoder on the fingerprints generated by the mass spectra encoder. Every bit that the encoder is unsure of leads to compounding errors when using the decoder.

Our core idea lies in choosing the right decoder that mitigates the weaknesses above. Since fingerprints represent the presence and absence of substructures in a molecule, we focus on capturing only present substructures, making it easier for the model to piece together the overall structure. Additionally, a decoder that uses the transformer architecture would be able to scale well with more training data. These two points can be met by using MOLFORGE [16] as the decoder. Our overall pipeline is shown in Figure 1.

Our key contributions are:

- We chose MIST [5] for spectrum-to-fingerprint encoding with MOLFORGE [16] decoder for our two-stage pipeline for *de novo* molecular structure generation from mass spectra.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat.

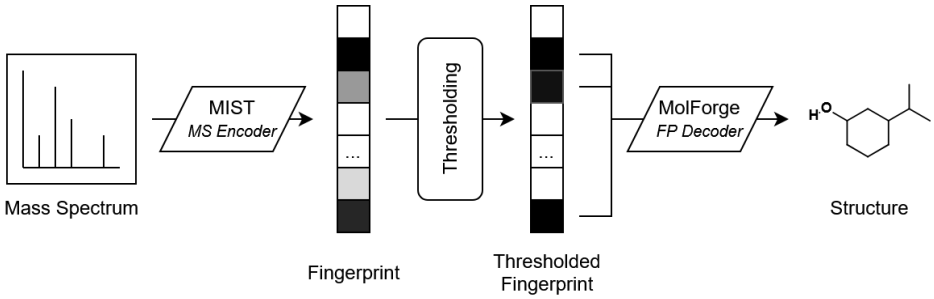

Figure 1: Our proposed pipeline of using MIST as a mass spectrum encoder, thresholding the fingerprint, and using MOLFORGE as a fingerprint decoder for the *de novo* molecule generation problem.

- We demonstrate that training the fingerprint-to-structure decoder on a larger dataset with varying chemical domains significantly enhances generalisation and structure recovery, even when fingerprint similarity to ground truth is moderate.
- Our method outperforms previous state-of-the-art by an order of magnitude, achieving 31% top-1 and 40% top-10 structure generation accuracy on the MassSpecGym [2] dataset.

2 Problem Statement & Definitions

We adopt the problem definition and metrics (top-k accuracy, Maximum Common Edge Subgraph (MCES) and Tanimoto similarity) for *de novo* molecule generation from MassSpecGym [2]. Further details can be found in Appendix A. Importantly, as this problem is *de novo* in nature, algorithms attempting to solve this problem need to generate molecular structures without any reference libraries.

3 Methodology

Our current method utilises a combination of existing models. We show how we have adapted these models to form a pipeline to solve the *de novo* molecule generation problem. These models (MIST, MOLFORGE) have open source implementations available. Details can be found in Appendix B.

3.1 Mass Spectra Encoder

We use the pretrained MIST models from DIFFMS. The MIST model was trained on MassSpecGym to predict molecular fingerprints from mass spectra, where each peak is annotated with possible chemical formulae by an automated annotation program. We refer the reader to Section 3.3 of the DIFFMS paper for technical details. In short, each peak in the mass spectra is encoded via a chemical formula transformer, then passed to a module that outputs the predicted fingerprint.

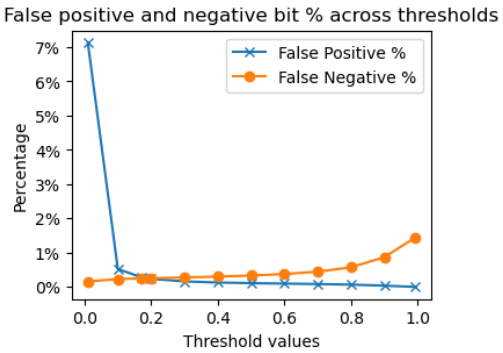

Figure 2: False positive and negative bits across different threshold values when applied to MIST fingerprints.

3.2 Fingerprint Decoder

MOLFORGE [16] is an autoregressive transformer model that predicts molecular structure from fingerprints. Specifically, it takes in the on-bits of a fingerprint (bits that have a value of '1') to predict a SMILES string that represents the molecular structure. This is done using an encoder-decoder model that generates a mapping of indices for the encoder and a mapping of tokens (representing atoms, parenthesis, and other symbols in SMILES strings) for the decoder.

MOLFORGE takes the indices of on-bits of the molecular fingerprint as input. However, MIST generates fingerprints as probabilities, real values between 0 to 1. As such, we set a threshold t and use all indices of bits $\geq t$ as input to MOLFORGE. We investigate two different threshold values: a fixed value $t = 0.5$, and in Section 4.4 a prior-adjusted threshold which assumes that the distribution

Table 1: *De novo* structural elucidation performance on MassSpecGym [2] dataset, across accuracy (Acc), MCES edit distance (MCES) and Tanimoto similarity of fingerprints (Tanimoto). For our pipeline (MIST+MolForge), we used a fixed threshold t of 0.5 and a prior-adjusted threshold value of 0.172, indicated by *.

	Top-1			Top-10		
Model	Acc. ↑	MCES ↓	Tanimoto ↑	Acc. ↑	MCES ↓	Tanimoto ↑
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
MIST+MolForge, t=0.5	28.27%	14.72	0.64	36.11%	10.69	0.70
***MIST+MolForge, t=0.172**	**30.97%**	**12.38**	**0.68**	**40.04%**	**8.63**	**0.74**

of on-bits in the training set is similar to that of the test set. Specifically, we choose t such that the proportion of logits thresholded to 1 is equal to the overall proportion of on-bits in the ground truth fingerprints of the training set, and we then apply this threshold to the test set logits. As shown in Figure 2, this prior-adjusted threshold can be calculated without any data leakage from the test set, and still results in a good balance of false positive and negative bits when compared to the ground truth fingerprints.

MassSpecGym only contains a limited number of molecules in the dataset ($\sim$17k), as there are multiple mass spectra for the same molecule. Thus, there is a need for further training such that MolForge has more data to learn the decoding process. We train MolForge with the same dataset used in DiffMS and its baselines, which contains $\sim$3M compounds and does not include any molecules in the test set. MolForge was trained to take in 4096-bit fingerprints as the pretrained MIST models generate fingerprints with a length of 4096 bits.

To generate the top-k molecular predictions, MolForge employs a beam search decoding strategy. At each decoding step, the model expands the most probable partial SMILES sequences by selecting the top-ranked token continuations according to the cumulative log-probability of each sequence. The search proceeds iteratively until complete SMILES strings are formed, after which the top-k highest probability sequences are retained as final outputs.

3.3 Datasets

We use MassSpecGym [2] to test our pipeline. For training on additional compounds, we refer to the combined dataset that is made available in DiffMS [1], which corresponds to data from DSSTox [3], HMDB [19], COCONUT [14], and MOSES [13] datasets. This results in about 2.8 million unique compounds, and we use this dataset as additional training for MolForge.

4 Results and Discussion

4.1 *De novo* molecule generation

The primary results of our model pipeline on the *de novo* molecular generation task are shown in Table 1. We benchmark our pipeline against the state-of-the-art DiffMS model, along with other established baselines reported in prior work. Our approach yields a significant performance gain, achieving an approximately 10-fold increase in exact structure match accuracy when MolForge is used as the fingerprint-to-structure decoder. Furthermore, the Tanimoto similarity between the predicted and ground truth molecular fingerprints show good agreement. The Maximum Common Edge Substructure (MCES) metric also shows a reduction in the number of extra atoms required to form a common graph between predicted and reference molecules, implying that predicted structures are more similar to the ground truth structure compared to other methods.

4.2 Performance of MIST as mass spectra encoder

As we have utilised the same pretrained MIST model from the DiffMS paper, we refer the reader to Section 4.4 of the DiffMS paper. DiffMS has also conducted an ablation study, showing that pretraining MIST results in improved performance compared to without pretraining MIST.

Table 2: Performance of MOLFORGE on different training datasets and the input fingerprint (FP) used. Additional training data leads to improved performance for MOLFORGE. We also compare ground truth fingerprints, generated directly from the ground truth structure, with fingerprints from MIST ($t = 0.172$), which are predicted from the corresponding mass spectra.

		Top-1		Top-10	
MolForge training dataset	Input FP	Acc. ↑	Tanimoto ↑	Acc. ↑	Tanimoto ↑
Combined DiffMS (∼3M)	From MIST	30.97%	0.68	40.04%	0.74
Combined DiffMS (∼3M)	Ground truth	46.00%	0.89	59.28%	0.93
MassSpecGym (∼17k)	Ground truth	0.00%	0.18	0.00%	0.19

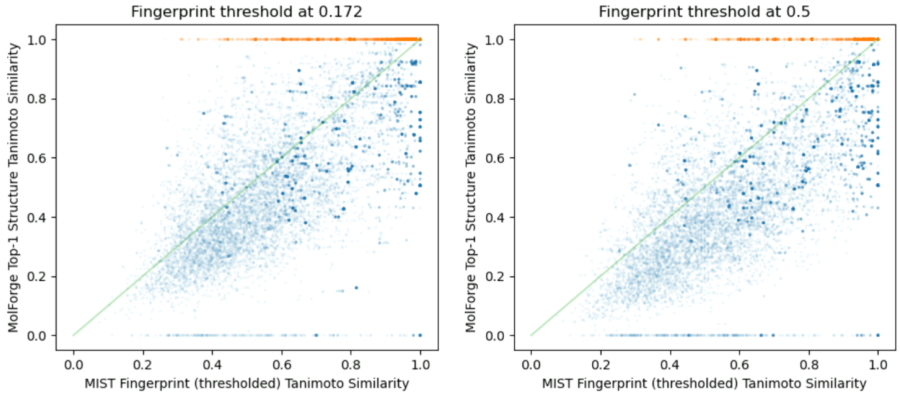

Figure 3: Tanimoto Similarity of thresholded MIST fingerprint and Top-1 structure from MOLFORGE, both compared to the ground truth fingerprint. Orange points indicate an exact structural match, and points above (below) the green parity line show predicted structures that are better (worse) than the thresholded fingerprint.

MIST also takes in annotations, in terms of chemical formulae, of peaks in the mass spectrum. In our study, removing annotations from the mass spectra (i.e. not supplying chemical formulae of the peaks into MIST) results in a decrease in Tanimoto Similarity (as calculated between fingerprint of the ground truth molecule and fingerprint predicted by MIST) from 0.731 to 0.627. Hence, peak annotations of the mass spectra are also essential for the good performance of our proposed pipeline.

4.3 Performance of MolForge as fingerprint decoder

We trained MOLFORGE with the same dataset used in DIFFMS and its baselines, which contains more than two million compounds. This allowed it to better capture the distribution of SMILES strings compared to just using the MassSpecGym dataset, which has only ∼17k compounds. Without these additional data, MOLFORGE fails to generate any structures from the test set correctly (see Table 2). The increase in performance with larger dataset size is expected due to data scaling laws [8]. Hence, we consider this step to be vital.

It is also important to note that the performance of MOLFORGE with ground truth fingerprints as input is very high with 59% top-10 accuracy (in terms of molecule structural matches) in the test set. This shows that MOLFORGE is a good model for the inverse problem of decoding fingerprints to structure, and why it is a vital component of our pipeline.

Additionally, instead of passing the probability of each bit in the fingerprint, thresholding the probabilities as a step function helps focus the decoder on the presence of substructures. MOLFORGE only takes in the indices of the fingerprint that has a value of 1, representing the presence of probable substructures in the fingerprint. This makes it capable of recovering accurate molecular structures even when the fingerprints predicted by MIST only moderately resembles the ground truth in terms of Tanimoto similarity, as shown in Figure 3. MOLFORGE predicts structures that has a better match to the ground truth in terms of Tanimoto similarity, compared to the thresholded fingerprint it was given as input, at 27% of the test set for $t = 0.5$ and 36% for a prior-adjusted threshold of $t = 0.172$.

Hence, the elevated performance of MOLFORGE with ground truth fingerprints shows that the key bottleneck in our pipeline is the prediction of fingerprints from mass spectra. The performance of our

pipeline is close to the best-case scenario if MIST were to predict fingerprints perfectly, suggesting room for improvement. Even though the fingerprint predicted by MIST has errors, MOLFORGE is able to overcome some of them and generate the correct structure. Further improvements on the prediction of fingerprints from mass spectra should translate well to better performance in the *de novo* molecule generation problem given the well-rounded performance of MOLFORGE.

4.4 Investigation on using prior-adjusted thresholds

The prior-adjusted threshold method calculates the prior distribution of on-bits in the training set and selects a threshold such that the proportion of thresholded logits matches this prior. For the MassSpecGym training set, the proportion of on-bits is 1.09%. To preserve this proportion, we determine the logit value corresponding to the top 1.09% of logits, which equals 0.172. Accordingly, we set the threshold to $t = 0.172$.

As shown in Table 1, using this prior-adjusted threshold results in a slight boost in performance, with the top-1 accuracy of 31% and top-10 accuracy of 40%. This is mainly attributed to the better balance of false positives and false negatives in the thresholded fingerprint. Hence, the indices passed to MOLFORGE would be more similar to the ground truth which results in the increased performance. We recommend using the prior-adjusted threshold value when using this pipeline.

5 Relevant Literature

***De novo* molecule generation from mass spectra**. The problem of *de novo* molecular generation from mass spectrometry data, also referred to as the inverse problem of structural elucidation from mass spectra, has garnered significant attention in computational chemistry and machine learning. Early approaches leveraged machine learning techniques to map mass spectra to molecular structures [11, 17]. The MassSpecGym benchmark [2] represents one of the most comprehensive open-source datasets for tandem mass spectrometry-based structure elucidation to compare across models. More recently, DIFFMS [1], a generative diffusion model, has been proposed to solve this problem. The method first employs MIST [5] to predict molecular fingerprints from mass spectra. These fingerprints are then used as conditional inputs to a graph diffusion model that generates candidate molecular structures as graphs. This two-stage framework enables DIFFMS to achieve a test-set accuracy of 2.30% on the MassSpecGym benchmark.

Encoding mass spectra to molecular fingerprints. DeepEI [7] also encodes a mass spectrum, but does not fully generate a molecular fingerprint. CSI:FingerID [4] uses a fragmentation tree to generate plausible fragments. However, this fingerprint type is different from Morgan fingerprints, and thus is not as extensible for usage with other fingerprint decoder models.

Decoding molecular fingerprints to molecular structure. DIFFMS [1], as stated earlier, contains a conditional graph diffusion model that takes in a fingerprint representation to predict molecule structure. Other fingerprint-to-structure models can be paired with MIST. Neuraldecipher [10] uses a standard feed-forward model to decode the molecular fingerprint, while MSNovelist [15] uses a recurrent neural network with long short-term memory architecture.

6 Conclusion

Various two-stage pipelines have been proposed for solving the *de novo* molecular generation problem from mass spectra, comprising (1) encoding mass spectra into molecular fingerprints, followed by (2) decoding these fingerprints into molecular structures. Leveraging external chemical datasets significantly enhances performance in both stages, especially so for the decoder. This stems from the limited availability of labeled mass spectra, in contrast to the abundance of molecular fingerprint data that can be exploited during decoder training. Notably, substituting the decoder with MOLFORGE, a seemingly modest architectural change, yields a tenfold improvement over prior state-of-the-art, highlighting the how choosing the right decoder is critical. MOLFORGE uses on-bit indices of the fingerprint as input, and we show how using a prior-adjusted threshold leads to further improvement compared to a predetermined value. We position this pipeline as a strong baseline for future work in *de novo* structure prediction from mass spectra, and highlight the spectra-to-fingerprint encoding step as a promising direction for further investigation.

References

[1] BOHDE, Montgomery ; MANJREKAR, Mrunali ; WANG, Runzhong ; JI, Shuiwang ; COLEY, Connor W.: *DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra*. Mai 2025

[2] BUSHUIEV, Roman ; BUSHUIEV, Anton ; DE JONGE, Niek F. ; YOUNG, Adamo ; KRETSCHMER, Fleming ; SAMUSEVICH, Raman ; HEIRMAN, Janne ; WANG, Fei ; ZHANG, Luke ; DÜHRKOP, Kai ; LUDWIG, Marcus ; HAUPT, Nils A. ; KALIA, Apurva ; BRUNGS, Corinna ; SCHMID, Robin ; GREINER, Russell ; WANG, Bo ; WISHART, David S. ; LIU, Li-Ping ; ROUSU, Juho ; BITTREMIEUX, Wout ; ROST, Hannes ; MAK, Tytus D. ; HASSOUN, Soha ; HUBER, Florian ; VAN DER HOOFT, Justin J. ; STRAVS, Michael A. ; BÖCKER, Sebastian ; SIVIC, Josef ; PLUSKAL, Tomáš: MassSpecGym: A Benchmark for the Discovery and Identification of Molecules. In: GLOBERSON, A. (Hrsg.) ; MACKEY, L. (Hrsg.) ; BELGRAVE, D. (Hrsg.) ; FAN, A. (Hrsg.) ; PAQUET, U. (Hrsg.) ; TOMCZAK, J. (Hrsg.) ; ZHANG, C. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 37, Curran Associates, Inc., 2024, S. 110010–110027

[3] CCTE, EPA: *Distributed Structure-Searchable Toxicity (DSSTox) Database*. 2024

[4] DÜHRKOP, Kai ; SHEN, Huibin ; MEUSEL, Marvin ; ROUSU, Juho ; BÖCKER, Sebastian: Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. In: *Proceedings of the National Academy of Sciences* 112 (2015), Oktober, Nr. 41, S. 12580–12585. – ISSN 0027-8424, 1091-6490

[5] GOLDMAN, Samuel ; WOHLWEND, Jeremy ; STRAŽAR, Martin ; HAROUSH, Guy ; XAVIER, Ramnik J. ; COLEY, Connor W.: Annotating Metabolite Mass Spectra with Domain-Inspired Chemical Formula Transformers. In: *Nature Machine Intelligence* 5 (2023), August, Nr. 9, S. 965–979. – ISSN 2522-5839

[6] HELLER, Stephen R. ; MCNAUGHT, Alan ; PLETNEV, Igor ; STEIN, Stephen ; TCHEKHOVSKOI, Dmitrii: InChI, the IUPAC International Chemical Identifier. In: *Journal of Cheminformatics* 7 (2015), Dezember, Nr. 1. – ISSN 1758-2946

[7] JI, Hongchao ; DENG, Hanzi ; LU, Hongmei ; ZHANG, Zhimin: Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks. In: *Analytical Chemistry* 92 (2020), Juli, Nr. 13, S. 8649–8653. – ISSN 1520-6882

[8] KAPLAN, Jared ; MCCANDLISH, Sam ; HENIGHAN, Tom ; BROWN, Tom B. ; CHESS, Benjamin ; CHILD, Rewon ; GRAY, Scott ; RADFORD, Alec ; WU, Jeffrey ; AMODEI, Dario: *Scaling Laws for Neural Language Models*. Januar 2020

[9] KRETSCHMER, Fleming ; SEIPP, Jan ; LUDWIG, Marcus ; KLAU, Gunnar W. ; BÖCKER, Sebastian: Coverage Bias in Small Molecule Machine Learning. In: *Nature Communications* 16 (2025), Januar, Nr. 1. – ISSN 2041-1723

[10] LE, Tuan ; WINTER, Robin ; NOÉ, Frank ; CLEVERT, Djork-Arné: Neuraldecipher – Reverse-Engineering Extended-Connectivity Fingerprints (ECFPs) to Their Molecular Structures. In: *Chemical Science* 11 (2020), Nr. 38, S. 10378–10389. – ISSN 2041-6520, 2041-6539

[11] LIM, Jing ; WONG, Joshua ; WONG, Minn X. ; TAN, Lee Han E. ; CHIEU, Hai L. ; CHOO, Davin ; NEO, Neng Kai N.: *Chemical Structure Elucidation from Mass Spectrometry by Matching Substructures*. November 2018

[12] MORGAN, H. L.: The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. In: *Journal of Chemical Documentation* 5 (1965), Mai, Nr. 2, S. 107–113. – ISSN 0021-9576, 1541-5732

[13] POLYKOVSKIY, Daniil ; ZHEBRAK, Alexander ; SANCHEZ-LENGELING, Benjamin ; GOLOVANOV, Sergey ; TATANOV, Oktai ; BELYAEV, Stanislav ; KURBANOV, Rauf ; ARTAMONOV, Aleksey ; ALADINSKIY, Vladimir ; VESELOV, Mark ; KADURIN, Artur ; JOHANSSON, Simon ; CHEN, Hongming ; NIKOLENKO, Sergey ; ASPURU-GUZIK, Alan ; ZHAVORONKOV, Alex: *Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models*. Oktober 2020

[14] SOROKINA, Maria ; MERSEBURGER, Peter ; RAJAN, Kohulan ; YIRIK, Mehmet A. ; STEINBECK, Christoph: COCONUT Online: Collection of Open Natural Products Database. In: *Journal of Cheminformatics* 13 (2021), Dezember, Nr. 1, S. 2. – ISSN 1758-2946

[15] STRAVS, Michael A. ; DÜHRKOP, Kai ; BÖCKER, Sebastian ; ZAMBONI, Nicola: MSNovelist: De Novo Structure Generation from Mass Spectra. In: *Nature Methods* 19 (2022), Juli, Nr. 7, S. 865–870. – ISSN 1548-7091, 1548-7105

[16] UCAK, Umit V. ; ASHYRMAMATOV, Islambek ; LEE, Juyong: Reconstruction of Lossless Molecular Representations from Fingerprints. In: *Journal of Cheminformatics* 15 (2023), Februar, Nr. 1. – ISSN 1758-2946

[17] WEI, Jennifer N. ; BELANGER, David ; ADAMS, Ryan P. ; SCULLEY, D: Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. In: *ACS central science* 5 (2019), Nr. 4, S. 700–708

[18] WEININGER, David: SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. In: *Journal of Chemical Information and Computer Sciences* 28 (1988), Februar, Nr. 1, S. 31–36. – ISSN 0095-2338, 1520-5142

[19] WISHART, David S. ; GUO, AnChi ; OLER, Eponine ; WANG, Fei ; ANJUM, Afia ; PETERS, Harrison ; DIZON, Raynard ; SAYEEDA, Zinat ; TIAN, Siyang ; LEE, Brian L. ; BERJANSKII, Mark ; MAH, Robert ; YAMAMOTO, Mai ; JOVEL, Juan ; TORRES-CALZADA, Claudia ; HIEBERT-GIESBRECHT, Mickel ; LUI, Vicki W. ; VARSHAVI, Dorna ; VARSHAVI, Dorsa ; ALLEN, Dana ; ARNDT, David ; KHETARPAL, Nitya ; SIVAKUMARAN, Aadhavya ; HARFORD, Karxena ; SANFORD, Selena ; YEE, Kristen ; CAO, Xuan ; BUDINSKI, Zachary ; LIIGAND, Jaanus ; ZHANG, Lun ; ZHENG, Jiamin ; MANDAL, Rupasri ; KARU, Naama ; DAMBROVA, Maija ; SCHIÖTH, Helgi B. ; GREINER, Russell ; GAUTAM, Vasuk: HMDB 5.0: The Human Metabolome Database for 2022. In: *Nucleic Acids Research* 50 (2022), Januar, Nr. D1, S. D622–D631. – ISSN 0305-1048, 1362-4962

A Problem Statement and Definitions

We adopt the problem definition for *de novo* molecule generation from MassSpecGym [2], which involves predicting a molecule's structure G given its corresponding mass spectrum MS. The molecule is modeled as a graph $G = (V, E)$, where the N atoms are represented by the set of vertices $V \in \mathbb{V}^N$, and the M chemical bonds are represented by the set of edges $E \in \mathbb{E}^M$. The mass spectrum MS consists of intensity values $Y \in (0, 1]$ associated with mass-to-charge ratios $X \in \mathbb{R}_+$.

This paper focuses on *de novo* molecule generation with the chemical formula provided as an additional input, which is also part of the MassSpecGym benchmark. In this setting, the chemical formula, which corresponds to the atom set V, is known in advance. The task is to predict the molecular structure G given its mass spectrum MS and chemical formula.

Importantly, as this problem is *de novo* in nature, algorithms attempting to solve this problem need to generate molecular structures without any reference libraries.

A.1 Metrics

Top-k metrics. The *de novo* molecule generation problem can be formulated as predicting a set of k candidate graphs $\widehat{G}_k = \{\widehat{G}_1, \ldots, \widehat{G}_k\}$, rather than a single predicted graph $\widehat{G}$. Especially since mass spectra may not contain enough information to predict structures accurately, this formulation better reflects the uncertainty and complexity of molecule generation from mass spectra.

To evaluate the quality of the predicted molecular graphs $\widehat{G}_k$, we compare them against the ground-truth graph G using three metrics, following MassSpecGym [2]. As a summary:

1. **Top-k accuracy.**

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

is the presence of the ground truth molecule within the top-k predictions of the model. This is averaged over all test examples. Here, $\mathbb{1}\{\cdot\}$ is the indicator function, which returns 1 if the condition is true and 0 otherwise.

2. **Maximum Common Edge Subgraph (MCES) metric.** [9]

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

The MCES metric measures the graph edit distance, or the number of edges that need to be added to candidate structure $\widehat{G}$ such that ground truth G is also a subgraph of it. We report the best similarity score among the top-k candidates as averaged across the test set. A score of 0 indicates identical graphs, while larger values correspond to greater structural dissimilarity.

3. **Tanimoto similarity**, computed on the Morgan fingerprints of the molecules [12].

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

This score reflects how well the generated molecule captures true molecular substructures. Tanimoto similarity ranges from 0 to 1, with 1 indicating perfect structural similarity (but not exact similarity, due to the possibility of fingerprint collisions).

A.2 Cheminformatics terms

While exact graph matching can be used to check if the ground truth and predicted molecules match, the cheminformatics domain has also developed other representations for this. In this paper, we focus on using InChIs [6] as textual representations of molecules. InChIs are guaranteed to be unique for each structure and thus can be used as a substitute to test for molecule similarity. SMILES strings [18], though a common text representation of molecules, are not used for comparison, as multiple SMILES strings can be used to refer to the same molecule.

We also use molecular fingerprints as a representation of a molecule, which are usually 2048 or 4096 bits in length. Molecular fingerprints are bit vectors that represent the presence ('1' / on-bit) or absence ('0' / off-bit) of substructures within a molecule. One of the most commonly used fingerprints is a Morgan fingerprint [12].

B Implementation Details

Open source implementations are available for the models (MIST [1], MolForge [2]) in our pipeline. We have generally followed the default settings for these models, and further modifications are explained in this section.

MIST: We use pretrained MIST models from DiffMS [3], and base our implementations from the DiffMS code and annotated MassSpecGym data. We proceed with the inference run with the MIST models pretrained for MassSpecGym, and collect the `output` of the `encoder` (MIST model) in `test_step` as the fingerprint outputs. We extracted all indices for which its corresponding bit is at least the threshold value.

MolForge: The training data was formatted following the examples in the MolForge Github repository, in which SMILES strings were split into tokens (i.e. atoms, brackets or numbers) and fingerprints were generated using the Morgan fingerprint generator from RDKit [4] with a radius of 2 and fingerprint size of 4096 bits. We first train the tokenizer for this model with the additional training data as provided in DiffMS. Following which, we trained the model on the same dataset with a learning rate of 5e-4 and batch size of 128 for 6 epochs. This took about 3 days to run with a Nvidia A40 GPU. For inference, we used the beam search code from MolForge with a beam size of 10.

[1] https://github.com/samgoldman97/mist/tree/main_v2

[2] https://github.com/knu-lcbc/MolForge

[3] https://zenodo.org/records/15122968

[4] https://www.rdkit.org