MS-BART: Unified Modeling of Mass Spectra and Molecules for Structure Elucidation

Yang Han[1,2], Pengyu Wang[1,2], Kai Yu[1,2,4], Xin Chen[2*], Lu Chen[1,2,3,4*]
[1]X-LANCE Lab, School of Computer Science
MoE Key Lab of Artificial Intelligence, SJTU AI Institute
Shanghai Jiao Tong University, Shanghai, China
[2]Suzhou Laboratory, Suzhou, China
[3]Shanghai Innovation Institute, Shanghai, China
[4]Jiangsu Key Lab of Language Computing, Suzhou, China
{csyanghan,chenlusz}@sjtu.edu.cn,mail.xinchen@gmail.com

Abstract

Mass spectrometry (MS) plays a critical role in molecular identification, significantly advancing scientific discovery. However, structure elucidation from MS data remains challenging due to the scarcity of annotated spectra. While large-scale pretraining has proven effective in addressing data scarcity in other domains, applying this paradigm to mass spectrometry is hindered by the complexity and heterogeneity of raw spectral signals. To address this, we propose MS-BART, a unified modeling framework that maps mass spectra and molecular structures into a shared token vocabulary, enabling cross-modal learning through large-scale pretraining on reliably computed fingerprint–molecule datasets. Multi-task pretraining objectives further enhance MS-BART's generalization by jointly optimizing denoising and translation task. The pretrained model is subsequently transferred to experimental spectra through finetuning on fingerprint predictions generated with MIST, a pre-trained spectral inference model, thereby enhancing robustness to real-world spectral variability. While finetuning alleviates the distributional difference, MS-BART still suffers molecular hallucination and requires further alignment. We therefore introduce a chemical feedback mechanism that guides the model toward generating molecules closer to the reference structure. Extensive evaluations demonstrate that MS-BART achieves SOTA performance across 5/12 key metrics on MassSpecGym and NPLIB1 and is faster by one order of magnitude than competing diffusion-based methods, while comprehensive ablation studies systematically validate the model's effectiveness and robustness. We provide the data and code at https://github.com/OpenDFM/MS-BART.

1 Introduction

Mass spectrometry is an analytical technique that measures the mass-to-charge ratio of ions, enabling the identification, quantification, and structural characterization of molecules. The identification of small molecules from mass spectrometry data represents a fundamental task in analytical chemistry, with broad applications across multiple domains, including drug discovery [1, 32, 39], environmental biochemistry [35], and materials science [21]. Recent advances in machine learning have enabled structure elucidation from mass spectra. Existing approaches can be broadly categorized into (1) retrieval-based methods and (2) *de novo* generative methods. Retrieval-based methods rely on matching query spectra against large annotated spectra databases. However, annotated experimental

* Xin Chen and Lu Chen are the corresponding authors.

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

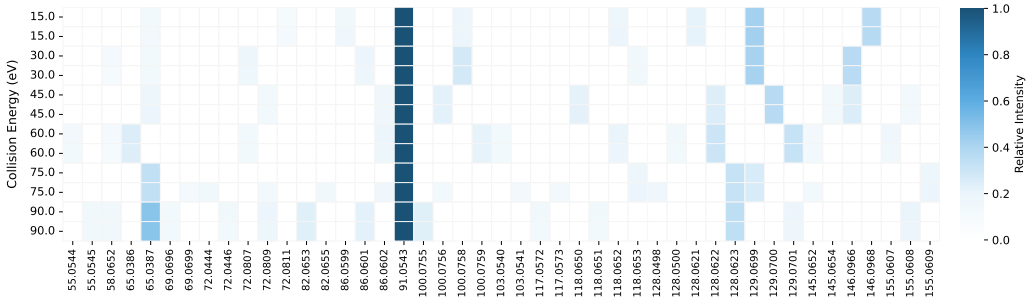

Figure 1: Randomly selected mass spectra of a molecule (SMILES: C#CCNCC1=CC=CC=C1, InChIKey: LDYBFSGEBHSTOQ) from MassSpecGym [7], acquired under varying collision energies. The x-axis shows the mass-to-charge ratio (m/z), the y-axis indicates collision energy (in eV), and color represents normalized relative intensity.

spectra are scarce and costly to obtain [3]. In addition, these methods are inherently limited to known molecules and cannot identify novel structures absent from reference databases. *De novo* generation methods learn to generate molecular structures directly from mass spectra, offering the potential to discover new compounds. However, these generative models [39, 45, 13, 29] are also bottlenecked by the limited availability of high-quality experimental spectra.

In other domains such as natural language processing (NLP) and computer vision (CV), a common strategy to overcome data scarcity is to pretrain models on large-scale unlabeled data and then finetune them on task-specific datasets [8, 28, 6, 20, 19]. A similar paradigm can also be observed in nuclear magnetic resonance (NMR) structure elucidation [48], where the model is first pretrained on 3.6 million unlabeled molecules and then fine-tuned directly on simulated and experimental NMR data. However, adapting this paradigm to mass spectrometry remains challenging due to the intrinsic complexity and heterogeneity of mass spectra. As illustrated in Fig. 1, spectra for the same molecule can vary substantially under different collision energies, adduct types, or instrument settings, and may even fluctuate slightly under identical experimental conditions. To address this variability, we propose using molecular fingerprints as an intermediate representation of mass spectra. Fingerprints are binary vectors that encode the presence of chemical substructures. Unlike raw spectra, they are invariant to experimental conditions and can be reliably computed from molecular structures using cheminformatics toolkits such as RDKit. This eliminates the need to simulate mass spectra under diverse experimental settings[34, 17], enabling scalable pretraining dataset construction.

Building on this insight, we propose MS-BART, a unified framework for molecular structure elucidation from mass spectrometry data, following the pretraining–finetuning–alignment paradigm widely used in NLP. We first construct a large-scale pretraining dataset consisting of fingerprint–molecule pairs, where molecular fingerprints are computed for 4 million unlabeled molecules using RDKit. Based on this dataset, we design multi-task pretraining objectives to facilitate cross-modal learning between molecular fingerprints and molecular structures. For finetuning, we incorporate experimental mass spectrometry data by using MIST [18] to predict fingerprints from spectra, conditioned on associated metadata and molecular formulas. As these predicted fingerprints are subject to dataset-specific noise and systematic biases, we finetune the pretrained model using the predicted fingerprints as input, thereby improving its adaptability to real-world experimental conditions. Finally, the generated structures are prone to molecular hallucinations [15], where outputs are chemically valid but deviate from the true molecules. To address this, we introduce an alignment step that incorporates chemical feedback by assigning higher probabilities to structures more similar to the ground truth. In summary, our main contributions are as follows:

- To the best of our knowledge, we are the first to leverage language model for mass spectra structure elucidation by introducing a unified vocabulary and multi-task pretraining on a large corpus of fingerprint–molecule pairs.
- We finetune the model on experimental data and incorporate chemical feedback to align the generative distribution with real-world structural preferences.

- We validate our approach on two public benchmarks, achieving SOTA performance across 5/12 key metrics on MassSpecGym[7] and NPLIB1[10] and is faster by one order of magnitude than competing diffusion-based methods.

2 Related Work

Mass Spectra Modeling. Mass spectra are variable-length, discrete, two-dimensional data, which makes them inherently challenging to model. A basic approach involves padding the two-dimensional matrix to a fixed length and projecting it into an embedding space via a linear layer [45]. A more common strategy is spectral binning, which partitions spectra into fixed-width intervals (e.g., 0.1 Da) to yield fixed-size input vectors [38]. For example, ChemEmbed [14] limits molecular weights to 700 Da, encodes spectra into a 7000-dimensional vector using a bin size of 0.01, and applies a convolutional neural network (CNN) to predict 300-dimensional Mol2vec embeddings. Similarly, Spec2Mol [29] and MS2DeepScore [23] convert spectra into bit vectors and train 1D CNNs or Siamese networks [5] to learn meaningful spectral embeddings. While binning is simple, it suffers from sparsity and sensitivity to noise, limiting its ability to capture chemically meaningful patterns. To overcome this, some methods leverage molecular fingerprints as intermediate representations that encode chemical substructures more robustly. CSI:FingerID [11] predicts molecular fingerprints from tandem mass spectra using fragmentation trees and machine learning, achieving strong performance in metabolite identification. MSNovelist [41] builds on this by integrating predicted fingerprints into an encoder-decoder model for de novo structure generation. Inspired by these approaches, MS-BART adopts molecular fingerprints as a spectrum representation, enabling scalable pretraining while preserving chemical semantics.

Structure Elucidation from Mass Spectra. Two major paradigms dominate structure elucidation from mass spectra: library matching and *de novo* generation. Library matching formulates the task as an information retrieval problem [40], comparing query spectra against databases of experimental or simulated spectra. Methods such as Spec2Vec [22] and MSBERT [49] learn spectrum embeddings and perform retrieval over databases like GNPS [44]. Due to the scarcity of experimental spectra, some methods (e.g., CFM-ID [43], GRAFF-MS [34]) simulate spectra from known molecular databases (e.g., PubChem). However, the effectiveness of library matching is constrained by database coverage, spectrum quality, and experimental variation, limiting its utility for novel compounds. In contrast, *de novo* approaches generate molecular structures directly from spectra, bypassing the need for reference databases. Spec2Mol [29] draws inspiration from speech-to-text models, employing an encoder-decoder network to translate spectra into SMILES sequences. MADGEN [45] uses a two-stage framework: scaffold retrieval and scaffold-conditioned molecule generation. Spectra and scaffolds are embedded into a shared latent space using MLPs and GNNs, and RetroBridge [24] generates the final structure conditioned on both inputs. Other models, including MSNovelist [41] and MS2SMILES [30], use fingerprints predicted by SIRIUS [9] as inputs to sequence models for SMILES generation. MS2SMILES further improves atom-level resolution by jointly predicting heavy atoms and their associated hydrogens. DiffMS [4] adopts an implicit fingerprint representation by extracting the final embedding from the precursor peak in MIST [18] and then generates the target structure by discrete diffusion conditioned on the spectrum embedding and node features (derived from the given formula). Despite promising results, most existing methods treat spectra and molecular structures as separate modalities, which often leads to semantic mismatches and molecular hallucinations [15]. In contrast, MS-BART unifies their modeling through a shared vocabulary. By pretraining on large-scale spectral fingerprint–molecule pairs, MS-BART learns rich representations for both chemical structures and their spectral abstractions. Subsequent finetuning and alignment on experimental spectra further enhance the model's ability to generate accurate and chemically consistent predictions on real-world data.

3 Methodology

Our framework is illustrated in Fig. 2. Section 3.1 introduces a unified vocabulary for representing both mass spectra and molecules. Section 3.2 describes multi-task pretraining with reliably computed fingerprints. Section 3.3 finetunes the model on experimental spectra to adapt to real-world distributions, while Section 3.4 further aligns molecular generation through chemical feedback.

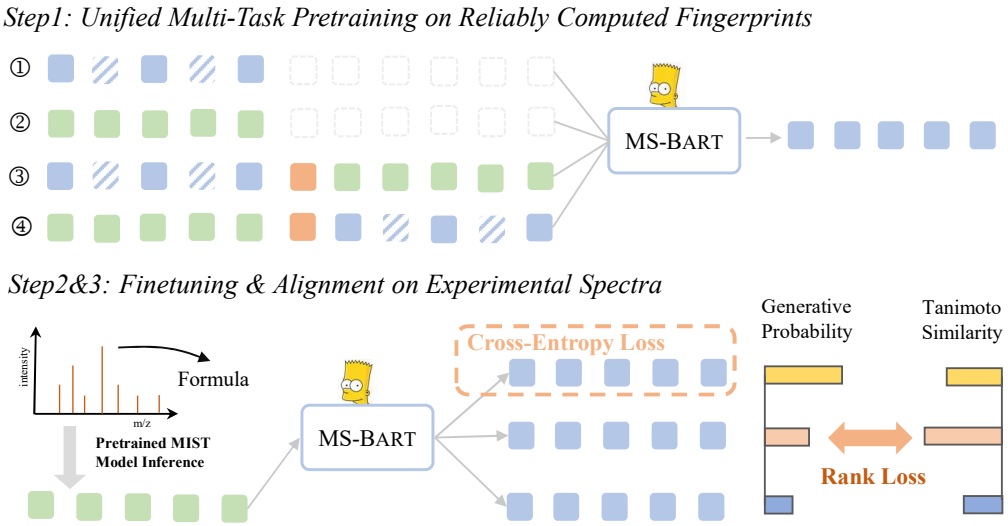

Figure 2: Overview of the MS-BART framework. Square symbols represent unified tokens, where ■ denotes SELFIES tokens, ■ indicates fingerprint tokens, and ■ represents a special separator token between SELFIES and fingerprint tokens. The ▢ symbol signifies padding tokens. Masked tokens are represented by striped patterns (▨). The top row illustrates pretraining tasks utilizing reliably computed fingerprints (Computed by RDKit), designed to enable MS-BART to capture fundamental patterns in both fingerprint and SELFIES representations. Transfer learning is subsequently applied to experimental spectra through finetuning and alignment to enhance structure elucidation performance.

3.1 Mass Spectra and Molecular Representation

Mass Spectra Representation. Raw mass spectra consist of variable-length sets of peaks $\{P_1, P_2, \ldots, P_k\}$, where each peak $P_i = (M_i, I_i)$ represents a mass-to-charge ratio (m/z) and a corresponding intensity. Due to noise and variability in experimental spectra, learning from raw spectra directly is challenging. In tandem mass spectrometry, *precursor ions* undergo collision-induced dissociation (CID), producing charged *product ions* and uncharged *neutral loss* fragments. These fragmentation patterns reflect structural information and correspond to chemical fragment encoded in molecular fingerprint bits. Prior work [11, 18, 2] has explored on predicting molecular fingerprints from spectra by leveraging fragmentation information. Following this, we represent each spectrum as a 4096-bit circular Morgan fingerprint $FP \in \{0,1\}^{4096}$, where each bit indicates the presence of a specific substructure. For experimental data, we employ the pretrained MIST model [18] to predict molecular fingerprints FP from mass spectra, conditioned on the chemical formula. The output of MIST is a probability vector $\{p_0, \ldots, p_{4095}\}$, where each p_i indicates the predicted likelihood of the i-th fingerprint bit being active. To convert these probabilities into a binary fingerprint representation, we apply a threshold ϵ:

$$FP_i = \begin{cases} 1, & \text{if } p_i \geq \epsilon, \\ 0, & \text{otherwise,} \end{cases} \quad \forall i \in \{0, 1, \ldots, 4095\}. \tag{1}$$

Each activated bit ($FP_i = 1$) is converted into a fingerprint token of the form `<fp{i:04d}>` (e.g., `<fp0123>`), producing a token sequence suitable for language modeling. We also compute the fingerprints from unlabeled molecules using RDKit [27] and apply the same tokenization for subsequent pretraining.

Molecular Representation. SMILES [46] and SELFIES [25] are two widely used string-based molecular representations. While SMILES is compact and human-readable, it does not guarantee chemical validity. In contrast, SELFIES is designed to ensure that every valid string maps to a chemically feasible molecule. We adopt SELFIES in MS-BART for its robustness and validity guarantees. To ensure consistency and uniqueness, we use the canonical form of each SELFIES string. Following [15], we employ a vocabulary of 185 SELFIES tokens. Although this is significantly smaller than typical language model vocabularies, prior work [37] shows that compact chemical vocabularies are sufficient for effective molecular representation learning.

3.2 Unified Multi-Task Pretraining on Reliably Computed Fingerprints

Given the tokenization approach described above, we tokenize mass spectra and molecular sequences by MIST [18] and the aforementioned vocabulary, respectively. As illustrated in Fig. 2, we design three self-supervised denoising tasks for pretraining to recover masked spans, along with one cross-modal translation task to strengthen modality alignment. We pretrain MS-BART on a simulated dataset of 4 million fingerprint–molecule pairs. The molecules are provided by MassSpecGym [7] and are preprocessed by excluding those with an MCES distance of less than two from any molecule in the test fold. The pretraining framework comprises four tasks detailed as follows: (1) *SELFIES Denoising (①)*. Randomly mask 30% of tokens in SELFIES sequence $S = \{s_1, \ldots, s_l\}$ with `[MASK]` token and recover original tokens. (2) *Fingerprint-to-Molecule Translation (②)*. Generate SELFIES sequences conditioned on fingerprint tokens FP. (3) *Hybrid Denoising (③ and ④)*. Combine fingerprint tokens and masked SELFIES using separator `<fps_sep>`, with input order variants, $[FP, \texttt{<fps\_sep>}, S_{masked}]$ and $[S_{masked}, \texttt{<fps\_sep>}, FP]$, predicting full SELFIES sequences from both modalities. All tasks follow a conditional generation paradigm optimized via cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{l} \log P(y_i \mid y_{<i}, X; \theta), \tag{2}$$

where θ denotes the model parameters, X is the input sequence (e.g., masked SELFIES or fingerprints), y_i is the i-th target token, and l is the target SELFIES length.

3.3 Finetuning on Experimental Spectra

After pretraining on the simulated fingerprint-molecule dataset, MS-BART is fine-tuned on experimental spectra to bridge the domain gap between computational and real-world data distributions. As shown in Fig. 2, the original mass spectra are tokenized into fingerprint tokens, and the model is also optimized through cross-entropy loss (Eq. 2) calculated between the target and predicted SELFIES tokens. This crucial step aims to learn the systematic bias introduced by the MIST [18] model and improve prediction performance.

3.4 Contrastive Alignment via Chemical Feedback

Following pretraining and dataset-specific fine-tuning, MS-BART acquires the capability to interpret molecular fingerprints and generate chemically plausible molecular structures. However, it remains susceptible to *molecular hallucination* [15], where generated molecules maintain chemical validity but exhibit limited consistency with original mass spectra or corresponding fingerprints. Specifically, the model may yield structures deviating substantially from true underlying molecular structures. To alleviate this hallucination and enhance performance, we propose aligning the model's probabilistic rankings of generated molecules with preference rankings derived from chemical contexts. In this paper, we define molecular preference through Tanimoto similarity, denoted as $Ps(\cdot)$. Given a fingerprint FP, MS-BART generates n candidate molecules $C = \{S_1, S_2, \cdots, S_n\}$. The preference score for each candidate S_i is calculated as $Ps(S_i) = Tan(S_i, S)$, where S represents the ground-truth molecular structure. Simultaneously, the model (parameterized by θ) assigns a conditional log-probability estimate $P_\theta(S_i)$ to each candidate S_i, given the input FP. Our objective is to establish consistency between the model's generative probabilities and Tanimoto similarity metrics. Specifically, for any candidate pair (S_i, S_j), we expect:

$$P_\theta(S_i) > P_\theta(S_j), if\ Ps(S_i) > Ps(S_j). \tag{3}$$

To encourage MS-BART to assign higher probabilities to candidate molecules that are more structurally similar to the target molecule, we employ a contrastive rank loss [31, 15], defined as:

$$\mathcal{L}_{\text{rank}}(C) = \sum_i \sum_{j>i} \max\left(0, P_\theta(S_j) - P_\theta(S_i) + \gamma_{ij}\right), \quad \forall i < j,\ \text{Ps}(S_i) > \text{Ps}(S_j), \tag{4}$$

where $\gamma_{ij} = (j - i) * \gamma$ denotes a margin scaled by the rank difference between candidates, γ is a hyperparameter. Additionally, we retain the token-level cross-entropy loss (Eq. 2) to preserve the

model's generative capability and the overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{\text{rank}}, \tag{5}$$

where α controls the weight of the rank loss. By jointly optimizing the token-level cross-entropy loss and the sequence-level contrastive rank loss on the same finetuning dataset, MS-BART can assign a balanced probability mass across the whole sequence. This optimization strategy elevates the probability of generating molecular structures that not only exhibit higher similarity to the target molecule but may also achieve exact matches.

4 Experiments

4.1 Datasets

We evaluate our MS-BART model on two widely used open-source benchmarks: NPLIB1 [10] and MassSpecGym [7], following prior works [4, 45]. NPLIB1 is a subset of the GNPS library, originally curated for training the CANOPUS model. Its name serves to distinguish the dataset from the associated tool. MassSpecGym is the largest publicly available dataset containing 231k high-quality mass spectra spanning 29k unique molecular structures. The dataset is partitioned into training, validation, and test sets based on the edit distance between molecular structures, facilitating robust evaluation. Although MADGEN [45] also reports performance on the NIST23 dataset, access to this resource is restricted due to its commercial licensing requirements.

4.2 Evaluation Metrics and Baselines

To evaluate the performance of our model, we employ the following metrics:

- **Top-k accuracy:** We measure the exact match between the predicted structure and the ground truth molecule by converting the generated molecule into a full InChIKey and comparing it with the gold InChIKey. Since MS fragmentation is largely insensitive to 3D stereochemistry, results based on 2D InChIKey are also reported in Appendix C.
- **Top-k maximum Tanimoto similarity:** This metric quantifies the structural similarity between molecules using molecular fingerprints. We compute fingerprints based on the Morgan algorithm [33] with a radius of 2 and a bit length of 2048 using RDKit.
- **Top-k minimum MCES (maximum common edge subgraph):** This metric measures the graph edit distance between molecules, reflecting the largest common substructure shared between the generated and ground-truth molecules [26].

We report the $k = 1, 10$ metrics following previous works [4, 7, 45]. Meanwhile, DIFFMS samples 100 molecules for each spectrum and identifies the top-k molecules based on frequency. To ensure a fair comparison, we sample 100 molecules and subsequently rank the generated molecules according to their distance from the given formula. Given two formula $F_1 = \{(a_1, n_1), (a_2, n_2), ..., (a_m, n_m)\}$ and $F_2 = \{(a_1, m_1), (a_2, m_2), ..., (a_m, m_k)\}$, the distance is defined as:

$$D(F_1, F_2) = \sum_{a \in \text{All Atoms}} |n_a - m_a|, \tag{6}$$

n_a, m_a are the counts of atom a in F_1 and F_2. If multiple molecules have the same distance, we sort them according to their estimated log-probability. After this re-ranking, we select the first k molecules as the top-k predictions.

Baselines. MassSpecGym [7] establishes three baselines for molecular generation: random generation, a SMILES-based Transformer, and a SELFIES-based Transformer. Spec2Mol [29] is retrained on both the NPLIB1 and MassSpecGym datasets to enable fair comparison. MIST+MSNovelist modifies the original MSNovelist framework [41] by replacing CSI:FingerID [11] with MIST. In MIST+Neuraldecipher, molecules are encoded into CDDD representations [47], followed by reconstruction of the SMILES strings using a pretrained LSTM decoder. MADGEN [45] and DIFFMS [4] represent recent state-of-the-art approaches. MADGEN first retrieves molecular scaffolds, then generates complete structures using the RetroBridge model [24], conditioned on both spectra and scaffolds. DIFFMS also adopts MIST as the spectrum encoder and employs a Graph Transformer [12] as the diffusion decoder, with separate pretraining of encoder and decoder components.

Table 1: Performance comparison of MS-BART and baseline methods on the NPLIB1 [10] and MassSpecGym [7]. Results marked with * are reproduced from MassSpecGym and DIFFMS. **Bold** denotes the best performance, underlined indicates the second-best.

Model	TOP-1			TOP-10		
	ACCURACY ↑	MCES ↓	TANIMOTO ↑	ACCURACY ↑	MCES ↓	TANIMOTO ↑
NPLIB1						
SPEC2MOL*	0.00%	27.82	0.12	0.00%	23.13	0.16
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
MADGEN	2.10%	20.56	0.22	2.39%	12.69	0.27
DIFFMS	**8.34%**	11.95	0.35	**15.44%**	9.23	0.47
MS-BART	7.45%	**9.66**	**0.44**	10.99%	**8.31**	**0.51**
MS-BART(Gold Fingerprint)	73.50%	2.14	0.90	79.12%	1.60	0.94
MASSSPECGYM						
SMILES TRANSFORMER*	0.00%	79.39	0.03	0.00%	52.13	0.10
SELFIES TRANSFORMER*	0.00%	38.88	0.08	0.00%	26.87	0.13
RANDOM GENERATION*	0.00%	21.11	0.08	0.00%	18.26	0.11
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
MIST + MSNOVELIST*	0.00%	45.55	0.06	0.00%	30.13	0.15
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**
MS-BART	1.07%	**16.47**	0.23	1.11%	15.12	0.28
MS-BART(Gold Fingerprint)	47.56%	3.26	0.85	64.62%	2.02	0.93

4.3 Implementation

We adopt BART-BASE [28] as the backbone of MS-BART, initializing all parameters from scratch using a normal distribution. To convert MIST probabilities into binary fingerprints, we apply a threshold of $\epsilon = 0.2$ for NPLIB1 and $\epsilon = 0.11$ for MassSpecGym. Further details regarding this selection are provided in Appendix A. During pretraining, we set the maximum sequence length to 512 to accommodate simultaneous learning from both fingerprints and SELFIES. For finetuning and alignment, we fix the input and output token lengths to 256, as the fingerprint inputs and SELFIES outputs rarely exceed this length in practice. When aligning MS-BART with chemical feedback, we freeze the encoder following practices from prior work [16, 36], and update only the decoder. Additional training details are provided in Appendix B.

4.4 Main Results

Table 1 presents a comparison of overall performance, demonstrating that MS-BART outperforms most baseline methods and achieves SOTA performance across 5/12 key metrics on NPLIB1 and MassSpecGym. Notably, on NPLIB1, MS-BART performs the best across all similarity metrics, surpassing the second-best method by 19.16% (MCES) and 25.71% (Tanimoto similarity) in the Top-1 setting. Significant improvements are also observed in the Top-10 setting, further validating the effectiveness of MS-BART. The high absolute Tanimoto similarity values (close to or exceeding 0.5) suggest that MS-BART generates structurally similar molecules, which are particularly valuable to domain experts. However, the Top-1 and Top-10 accuracy do not surpass DiffMS [4], primarily because we have filtered out similar molecules in the pretraining data (Appendix B), whereas DiffMS only removes all NPLIB1 and MassSpecGym test and validation molecules from their pretraining dataset. Another recently proposed open-source dataset, MassSpecGym, presents a more challenging benchmark than NPLIB1, as NPLIB1 lacks a scaffold-based split, resulting in a test set containing molecules with high structural similarity (Tanimoto similarity > 0.85) to those in the training set [4, 7]. Meanwhile, MassSpecGym exhibits a more complex data composition due to the presence of $[M+Na]^+$ adducts. The Na^+ atom have a higher mass than H^+ atom, leading to more complex fragmentation patterns and consequently a data distribution that differs significantly from that of $[M+H]^+$. Furthermore, the $[M+Na]^+$ data in the MassSpecGym training set is relatively scarce, constituting only 15.52% of the total samples. This class imbalance, combined with the distinct fragmentation behavior of $[M+Na]^+$ compared to $[M+H]^+$, introduces additional noise and potential bias during model training. To avoid fragmentation pattern conflicts and preserve data consistency, we

Table 2: Performance comparison of MS-BART on the NPLIB1 dataset using different pretraining strategies. "NONE" indicates no pretraining, "SD" and "TRANS" denote pretraining with *SELFIES Denoising* and *Fingerprint-to-Molecule Translation*, respectively, and "HYBRID" refers to pretraining with the *Hybrid Denoising* method described in Section 3.2.

PRETRAIN STRATEGY	TOP-1			TOP-10		
	ACCURACY ↑	MCES ↓	TANIMOTO ↑	ACCURACY ↑	MCES ↓	TANIMOTO ↑
NONE	1.71%	12.93	0.27	3.05%	11.36	0.34
SD	0.37%	14.41	0.24	0.98%	12.42	0.32
TRANS	6.23%	**9.37**	0.42	10.26%	**7.98**	0.50
HYBRID	5.13%	9.96	0.41	7.81%	8.87	0.48
MS-BART	**7.45%**	9.66	**0.44**	**10.99%**	8.31	**0.51**

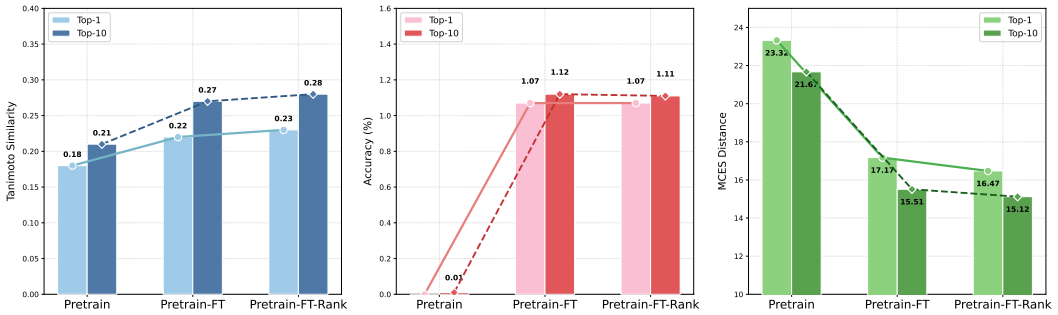

Figure 3: Progressive improvement of MS-BART for molecular hallucination mitigation on MassSpecGym. Subfigures show: a) Tanimoto similarity, b) Accuracy, and c) MCES scores across three training stages under Top-1 and Top-10 settings. Stage 1: Pretrain, the base model pretrained on 4M simulated unlabeld molecules. Stage 2: Pretrain-FT, fine-tuned on MassSpecGym. Stage 3: Pretrain-FT-Rank, fully optimized MS-BART with chemical feedback.

filter out $[M+Na]^+$ adducts before fine-tuning and alignment, retaining only the dominant $[M+H]^+$ data. However, to ensure a fair and comprehensive evaluation, we retain the $[M+Na]^+$ adducts in the test set. The results show that MS-BART does not surpass DiffMS on most metrics, and the main reason is same as for NPLIB1. Excluding DiffMS, MS-BART also shows SOTA performance across all similarity metrics and demonstrates the robustness and effectiveness of MS-BART in handling diverse spectral patterns. Finally, we report the best possible performance of MS-BART if we use the gold fingerprint calculated from the true structure rather than predicting it with the pretrained MIST model. It is noteworthy that MS-BART can almost find the exact match or extremely similar candidates, proving the great potential of MS-BART and indicating that further work can be devoted to improving the performance of MIST model.

4.5 Unified Multi-Task Pretraining Enhances Cross-Modal Learning

Pretraining serves as the foundational phase and a critical step in training language models, enabling a broad understanding of linguistic features such as syntax, semantics, and context, which are essential for effective transfer learning. To investigate the role of multi-task pretraining in enhancing MS-BART's comprehension of molecular fingerprints and SELFIES, we conduct ablation experiments, with results presented in Table 2. It is obvious that the model trained without pretraining exhibits relatively poor performance compared to its pretrained counterpart. Nevertheless, its accuracy remains above chance level and surpasses baseline methods that encode mass spectra directly, demonstrating the advantage of representing raw mass spectra as fingerprints and training with a unified vocabulary in an end-to-end style. Furthermore, we compare multi-task pretraining with single-task pretraining. Pretraining solely with the denoising task leads to performance degradation rather than improvement, primarily because the denoising task is not well aligned with structure elucidation. The substantial improvement observed in the fingerprint-to-molecule translation task further supports this finding. Moreover, the performance gain of MS-BART over single-task pretraining indicates that the denoising task remains beneficial, as it helps MS-BART develop a fundamental understanding of molecular

structures and contributes to the final performance. These results suggest that unified multi-task pretraining on unlabeled data promotes cross-modal interaction and alignment by enabling MS-BART to learn a shared representation space across modalities.

4.6 MS-BART Mitigates Molecular Hallucination

The training paradigm of MS-BART consists of three steps. Fig. 3 illustrates the impact of these steps on the final performance in MassSpecGym, revealing two key findings. First, continued fine-tuning on the MassSpecGym training fold improves performance and mitigates molecular hallucination, particularly in terms of accuracy. For example, the "Pretrain" model achieves near-zero Top-1 and Top-10 accuracy scores of 0.00% and 0.01%, respectively, while the "Pretrain-FT" model improves these values to 1.07% and 1.12%. Moreover, aligning the model with chemical feedback based on Tanimoto similarity further reduces molecular hallucination. This alignment enables the generation of structurally more similar molecules, as evidenced by the improved Tanimoto similarity and the degraded MCES metric. Fig. 4 shows a randomly selected case from MassSpecGym where MS-BART effectively mitigates molecular hallucination.

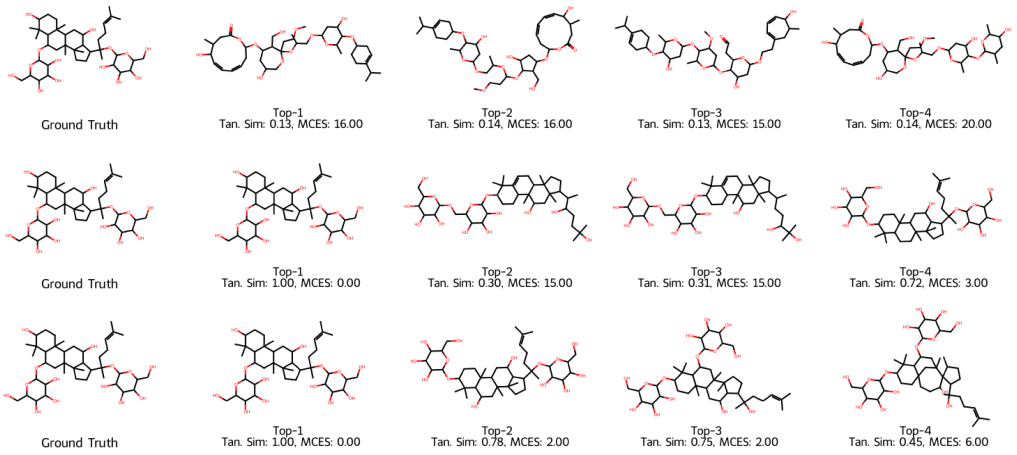

Figure 4: Predictions of the Pretrain (top row), Pretrain-FT (middle row), and Pretrain-FT-Rank (bottom row) models on a representative MassSpecGym sample, with the ground-truth structure shown in the first column and the Top-4 generated outputs in the subsequent columns. Results demonstrate that MS-BART can effectively mitigate molecular hallucination and predict more chemically plausible and structurally consistent predictions.

4.7 Sensitivity Analysis of Model Hyperparameters

MS-BART employs two key hyperparameters, the fingerprint generation threshold ϵ and the rank loss weight α. The fingerprint threshold determines how the MIST probability is converted into an active fingerprint, where different thresholds lead to distinct representations. The rank loss weight controls the influence of the token-level rank loss in guiding the alignment training. To comprehensively evaluate the impact of these two parameters, we experiment with different values of ϵ on MassSpecGym and α on NPLIB1 and present the results in Table 3. The results show no significant differences among the tested ϵ values, indicating that MS-BART is not sensitive to this parameter. Although $\epsilon = 0.11$ achieves the best performance on the MassSpecGym validation set and is reported in Table 1 as the final result, the ablation results suggest that $\epsilon = 0.11$ is not the best overall. This discrepancy is mainly due to the high difficulty of the MassSpecGym dataset and the distribution mismatch between its validation and test sets. A similar observation holds for α on NPLIB1. The model with $\alpha = 5$ performs best in terms of Top-1 Tanimoto score on the validation set and is also reported as the final result. However, models with $\alpha = 1$ and $\alpha = 3$ also achieve competitive results on specific metrics. The variations across different configurations are minor, indicating that MS-BART is not highly sensitive to the rank loss weight.

4.8 Analysis of Decoding Hyperparameters

Table 3: Performance comparison of MS-BART under two key model hyperparameters, where the ablation of the fingerprint generation threshold and the rank loss weight are evaluated on MassSpecGym and NPLIB1, respectively.

	TOP-1			TOP-10		
	ACCURACY ↑	MCES ↓	TANIMOTO ↑	ACCURACY ↑	MCES ↓	TANIMOTO ↑
Fingerprint Generation Threshhold ϵ (MassSpecGym)						
$\epsilon = 0.11$	1.07%	16.47	0.23	1.11%	15.12	0.28
$\epsilon = 0.15$	1.20%	16.88	0.23	1.20%	15.45	0.28
$\epsilon = 0.20$	1.19%	17.27	0.23	1.20%	15.75	0.28
Rank Loss Weight α (NPLIB1)						
$\alpha = 1$	7.08%	10.56	0.44	12.45%	9.10	0.52
$\alpha = 3$	7.20%	9.53	0.44	10.50%	8.21	0.51
$\alpha = 5$	7.45%	9.66	0.44	10.99%	8.31	0.51

MS-BART employs beam-search multinomial sampling for structure prediction, governed by two key decoding hyperparameters while maintaining default values for other parameters. First, the temperature parameter [50, 42] regulates output randomness during inference, a factor known to substantially influence generation quality in typical language models. Interestingly, MS-BART's outputs demonstrate remarkable stability across temperature variations, which is likely attributable to the model's high confidence in next-token prediction gained through multi-task pretraining and its relatively small vocabulary size of 185. Detailed temperature analysis is provided in Appendix D. The second critical parameter, beam width, governs the search space breadth during decoding. We systematically evaluated MS-BART's performance on the complete MassSpecGym test fold using an NVIDIA A800-SXM4-80GB GPU, exploring beam widths from 10 to 100. As shown in Fig. 5, both Top-1 and Top-10 accuracies demonstrate consistent improvement with increasing beam width, while inference latency exhibits linear scaling. Notably, when tested on a common consumer-grade GPU like the RTX 4090 with a beam width of 100, the average inference time per spectrum is about 3 seconds, which is 53x times faster than DiffMS's approximately 160 seconds and remains practically acceptable.

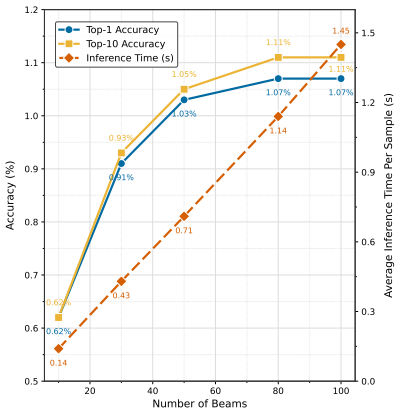

Figure 5: Performance of MS-BART with beam-search multinomial sampling decoding under different beam widths: Top-1 accuracy, Top-10 accuracy, and average inference time per spectrum.

5 Conclusion

In this work, we propose MS-BART, a novel language model for mass spectra structure elucidation within a unified modeling framework. Specifically, we first represent mass spectra as fingerprints and tokenize both the fingerprints and molecular representations (SELFIES) using a unified vocabulary. Subsequently, MS-BART undergoes the standard pretraining-finetuning-alignment paradigm commonly employed in NLP, enabling the model to interpret spectral fingerprints and generate plausible molecular structures. Extensive experiments demonstrate that MS-BART achieves SOTA performance on two widely adopted benchmarks across 5/12 key metrics on MassSpecGym and NPLIB1 and is faster by one order of magnitude than competing diffusion-based method, while ablation studies further validate its effectiveness and robustness.

Acknowledgments and Disclosure of Funding

This work was supported by the National Science and Technology Major Project (2023ZD0120703), the China NSFC Projects (U23B2057, 62120106006, and 92370206), and Shanghai Municipal Science and Technology Projects (2021SHZDZX0102 and 25X010202846).

References

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[2] Sadjad Fakouri Baygi and Dinesh Kumar Barupal. Idsl_mint: a deep learning framework to predict molecular fingerprints from mass spectra. *Journal of Cheminformatics*, 16(1):8, 2024.

[3] Wout Bittremieux, Mingxun Wang, and Pieter C Dorrestein. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics*, 18(12):94, 2022.

[4] Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W Coley. Diffms: Diffusion generation of molecules conditioned on mass spectra. *arXiv preprint arXiv:2502.09571*, 2025.

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Roman Bushuiev, Anton Bushuiev, Niek de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[9] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A Aksenov, Alexey V Melnik, Marvin Meusel, Pieter C Dorrestein, Juho Rousu, and Sebastian Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods*, 16(4):299–302, 2019.

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

[11] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.

[12] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

[13] David Elser, Florian Huber, and Emmanuel Gaquerel. Mass2smiles: deep learning based fast prediction of structures and functional groups directly from high-resolution ms/ms spectra. *bioRxiv*, pages 2023–07, 2023.

[14] Muhammad Faizan-Khan, Roger Giné, Josep M Badia, Maribel Pérez-Ribera, Alexandra Junza, Maria Vinaixa, Marta Sales-Pardo, Roger Guimerà, and Oscar Yanes. Chemembed: A deep learning framework for metabolite identification using enhanced ms/ms data and multidimensional molecular embeddings. *bioRxiv*, pages 2025–02, 2025.

[15] Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. Domain-agnostic molecular generation with chemical feedback. In *The Twelfth International Conference on Learning Representations*.

[16] Giacomo Frisoni, Paolo Italiani, Stefano Salvatori, and Gianluca Moro. Cogito ergo summ: abstractive summarization of biomedical papers via semantic parsing graphs and consistency rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12781–12789, 2023.

[17] Samuel Goldman, John Bradshaw, Jiayi Xin, and Connor Coley. Prefix-tree decoding for predicting mass spectra from molecules. *Advances in neural information processing systems*, 36:48548–48572, 2023.

[18] Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J Xavier, and Connor W Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, 2023.

[19] Yang Han, Yiming Wang, Rui Wang, Lu Chen, and Kai Yu. Alignsum: Data pyramid hierarchical fine-tuning for aligning with human summarization preference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8506–8522, 2024.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[21] Xiu Huang, Huihui Liu, Dawei Lu, Yue Lin, Jingfu Liu, Qian Liu, Zongxiu Nie, and Guibin Jiang. Mass spectrometry for multi-dimensional characterization of natural and synthetic materials at the nanoscale. *Chemical Society Reviews*, 50(8):5243–5280, 2021.

[22] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.*, 17(2), 2021.

[23] Florian Huber, Sven van der Burg, Justin JJ van der Hooft, and Lars Ridder. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics*, 13(1):84, 2021.

[24] Ilia Igashov, Arne Schneuing, Marwin Segler, Michael M Bronstein, and Bruno Correia. Retrobridge: Modeling retrosynthesis with markov bridges. In *The Twelfth International Conference on Learning Representations*.

[25] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

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

[27] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[28] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871. Association for Computational Linguistics, 2020.

[29] Eleni E Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia E Kavraki. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry*, 6(1):132, 2023.

[30] Yanmin Liu, Xuan Zhang, Wei Zhao, Daming Zhu, and Xuefeng Cui. De novo molecular structure generation from mass spectra. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 373–378. IEEE, 2023.

[31] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. BRIO: Bringing order to abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[32] Felix Meissner, Jennifer Geddes-McAlister, Matthias Mann, and Marcus Bantscheff. The emerging role of mass spectrometry-based proteomics in drug discovery. *Nature Reviews Drug Discovery*, 21(9):637–654, 2022.

[33] Harry L Morgan. 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.

[34] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks. In *International Conference on Machine Learning*, pages 25549–25562. PMLR, 2023.

[35] Yolanda Picó and Damià Barceló. Pyrolysis gas chromatography-mass spectrometry in environmental analysis: Focus on organic matter and microplastics. *TrAC Trends in Analytical Chemistry*, 130:115964, 2020.

[36] Alessandro Raganato and Jörg Tiedemann. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 287–297, 2018.

[37] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[38] Khawla Seddiki, Philippe Saudemont, Frédéric Precioso, Nina Ogrinc, Maxence Wisztorski, Michel Salzet, Isabelle Fournier, and Arnaud Droit. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nature communications*, 11(1):5595, 2020.

[39] Michael A Skinnider, Fei Wang, Daniel Pasin, Russell Greiner, Leonard J Foster, Petur W Dalsgaard, and David S Wishart. A deep generative model enables automated structure elucidation of novel psychoactive substances. *Nature Machine Intelligence*, 3(11):973–984, 2021.

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

[41] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. Msnovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022.

[42] Chi Wang, Xueqing Liu, and Ahmed Hassan Awadallah. Cost-effective hyperparameter optimization for large language model generation inference. In Aleksandra Faust, Roman Garnett, Colin White, Frank Hutter, and Jacob R. Gardner, editors, *Proceedings of the Second International Conference on Automated Machine Learning*, volume 224 of *Proceedings of Machine Learning Research*, pages 21/1–17. PMLR, 12–15 Nov 2023.

[43] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S Wishart. Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analytical chemistry*, 93(34):11692–11700, 2021.

[44] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016.

[45] Yinkai Wang, Xiaohui Chen, Liping Liu, and Soha Hassoun. MADGEN: Mass-spec attends to de novo molecular generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[46] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[47] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

[48] Lin Yao, Minjian Yang, Jianfei Song, Zhuo Yang, Hanyu Sun, Hui Shi, Xue Liu, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Conditional molecular generation net enables automated structure elucidation based on 13c nmr spectra and prior knowledge. *Analytical chemistry*, 95(12):5393–5401, 2023.

[49] Hailiang Zhang, Qiong Yang, Ting Xie, Yue Wang, Zhimin Zhang, and Hongmei Lu. Msbert: Embedding tandem mass spectra into chemically rational space by mask learning and contrastive learning. *Analytical Chemistry*, 96(42):16599–16608, 2024.

[50] Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 437–445, 2024.

A Selection of the Probability Threshold ϵ

We selected the threshold value ϵ for NPLIB1 and MassSpecGym using different methods. Since NPLIB1 is known to be a much easier dataset, we chose its threshold empirically. After examining the probability distribution, we observed that most probabilities were smaller than 0.3. A large threshold could cause information loss, whereas a small threshold might introduce noise. Therefore, we selected $\epsilon = 0.2$ to balance these factors, which resulted in excellent performance. For MassSpecGym, we trained MS-BARTusing ϵ values ranging from 0.1 to 0.2 in increments of 0.01. We then validated these on the validation set and selected the best threshold based on the highest Top-1 Tanimoto similarity. As shown in Table 4, the Top-1 Tanimoto similarity values for different thresholds were very close. We ultimately chose $\epsilon = 0.11$ as the final threshold, although performance did not vary significantly across the different threshold values.

Table 4: Top-1 Tanimoto Similarity under different ϵ values.

ϵ	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
Tanimoto Similarity	0.1666	**0.1678**	0.1640	0.1651	0.1660	0.1660	0.1654	0.1643	0.1651	0.1649	0.1636

B MS-BART Training Details

Pretraining. We pretrain MS-BART from scratch on a simulated dataset of 4M fingerprint-molecule pairs. This dataset was generated from a refined subset of the 4M unlabeled molecules provided by MassSpecGym [7]. The refinement process involved excluding any molecule with an MCES distance of less than two from any molecule in the MassSpecGym test fold. For the NPLIB1 dataset, we first evaluated the structural similarity between its test set and our pretraining data. As shown in Table 5, approximately 3% of the pretraining molecules were structurally similar to those in the NPLIB1 test set. To prevent data leakage, we further filtered the pretraining set to remove any molecules that were structurally similar or identical to those in the NPLIB1 test set. Specifically, we refined the

pretraining dataset by removing molecules with a maximum Tanimoto similarity greater than 0.5 to any molecule in the NPLIB1 test set.

The training was conducted on four NVIDIA A800-SXM4-80GB GPUs using bfloat16 precision. A per-device batch size of 96 and two gradient accumulation steps were employed, resulting in an effective total batch size of 768 across three training epochs. The optimization process adopted a cosine learning rate scheduler with a warm-up phase of 10,000 steps. The learning rate increased linearly from zero to a peak value of 6e-4 during warm-up and subsequently decayed following a cosine schedule to a minimum value of 1e-5. The entire multi-task pretraining process required approximately 34 hours to complete.

Table 5: Distribution of max Tanimoto Similarity between pretraining set and NPLIB1 test set.

Similarity Interval	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Proportion	0.09%	8.40%	55.85%	21.58%	6.79%	3.79%	1.84%	1.20%	0.40%	0.06%

Finetuning. Fine-tuning is performed on a single NVIDIA A800-SXM4-80GB GPU using bfloat16 precision, with consistent parameter settings across both MassSpecGym and NPLIB1 datasets. We adopt a learning rate of 5e-5 combined with a warm-up phase covering 10% of total training steps. Each training iteration processes 128 samples per batch. For validation monitoring, we implement an early stopping criterion that evaluates model performance every 400 steps on MassSpecGym and every 200 steps on NPLIB1. Training terminates when the validation set's Top-1 Tanimoto similarity fails to improve for three consecutive evaluations.

Alignment Training. The alignment phase is more complex than pretraining and fine-tuning, requiring careful hyperparameter optimization. As specified in Table 6, we explore multiple configurations during this stage while maintaining a fixed batch size of 128 and bfloat16 precision on a single NVIDIA A800 GPU. The learning schedule includes a 10% warm-up ratio of total training steps. Validation frequency is set to 400-step intervals for MassSpecGym and 50-step intervals for NPLIB1. We extend the patience to five consecutive evaluations without improvement in Top-1 Tanimoto similarity before triggering early stopping. Final hyperparameter selection is determined by maximizing the validation set's Top-1 Tanimoto similarity.

Table 6: Hyper-parameter settings.

Hyper-parameters	Values
Learning Rate	{1e-5, 5e-5}
Candidate Margin γ	{0.05, 0.1, 0.2}
Rank Loss Weight α	{1, 3, 5}
Number of Candidates	{3, 5}
Length Penalty Coefficient	{1.4, 1.6}

C 2D InChIKey Based Accuracy

We evaluate the Top-1 and Top-10 accuracy based on 2D InChIKey matching and present the results in Table 7. On NPLIB1, the two calculation methods yield the same results. However, for MassSpecGym, the 2D InChIKey-based accuracy of MS-BART shows a slight improvement compared to evaluation with the full InChIKey because the 2D InChIKey uses only the first 14 characters, which do not include 3D stereochemistry. To maintain consistency with the DiffMS [4], we use the full InChIKey results as the final results.

D The Impact of Sampling Temperature on Model Performance

Table 8 presents the performance evaluation of MS-BART during decoding with varying temperature values on MassSpecGym. The results demonstrate that MS-BART's performance remains largely consistent across different temperature settings. This stability can be attributed to the model's high confidence in next-token predictions, which likely stems from its multi-task pretraining framework.

Table 7: 2D InChIKey based Top-1 and Top-10 accuracy on NPLIB1 and MassSpecGym.

Calculation Method	Top-1 Accuracy	Top-10 Accuracy
NPLIB1		
MS-BART (InChIKey)	7.45%	10.99%
MS-BART (2D InChIKey)	7.45%	10.99%
MASSSPECGYM		
MS-BART (InChIKey)	1.07%	1.11%
MS-BART (2D InChIKey)	1.26%	1.28%

Table 8: The performance of MS-BART when decoding with different temperature on MassSpecGym.

TEMPERATURE	TOP-1			TOP-10		
	ACCURACY ↑	MCES ↓	TANIMOTO ↑	ACCURACY ↑	MCES ↓	TANIMOTO ↑
Prtrain-FT						
0.2	1.07%	17.16	0.22	1.12%	15.51	0.27
0.4	1.07%	17.17	0.22	1.12%	15.51	0.27
0.8	1.07%	17.16	0.22	1.12%	15.50	0.27
MS-BART						
0.2	1.07%	16.47	0.23	1.11%	15.12	0.28
0.4	1.07%	16.47	0.23	1.11%	15.12	0.28
0.8	1.07%	16.47	0.23	1.11%	15.11	0.28

E Limitations

Our primary limitation lies in the reliance on the external MIST model [18] for predicting fingerprints from experimental spectra. The accuracy of these fingerprint predictions significantly impacts the overall performance. Future work may focus on fine-tuning the MIST model using task-specific datasets to improve prediction accuracy, or it could explore directly modeling the fragments from raw mass spectra. Another limitation is that the formula information is used only for re-ranking and not for training the model. However, these additional formulas definitely include important information and are very likely to boost the model's performance. We plan to explore effective ways to incorporate these informational hints into MS-BART in future work.

F Ethics Statement

All data used in this work are obtained from publicly available sources, including molecular structure datasets and mass spectrometry benchmarks such as NPLIB1 and MassSpecGym. We strictly follow the corresponding licenses and usage protocols. No modifications have been made to the original datasets beyond necessary preprocessing using standard cheminformatics tools (e.g., RDKit) to compute molecular fingerprints. No personal, private, or sensitive data are involved. The proposed model, MS-BART, is trained and evaluated solely for the task of molecular structure elucidation from mass spectrometry data. Our work does not involve human subjects, biometric data, or decision-making in socially sensitive applications. We do not foresee any immediate negative societal impact. On the contrary, improving molecular identification from mass spectrometry may benefit fields such as drug discovery, environmental chemistry, and materials science.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [Abstract, Section 1]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: [Appendix E]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [The paper does not include theoretical results]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: [Section 4.14.3, Appendix B]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [Section 4.1, Anonymous code]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: [Section 4.14.3, Appendix B]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: [Appendix D]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [Section 4.8, Appendix B]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: [Appendix F]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [Appendix F]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: [Appendix F]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: [Section 4.1]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [We provide the model parameters.]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [Our paper does not involve crowdsourcing experiments and research.]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [We use LLM for writing.]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.