Breaking the Modality Barrier: Generative Modeling for Accurate Molecule Retrieval from Mass Spectra

Yiwen Zhang[1,2], Keyan Ding[2*], Yihang Wu[1], Xiang Zhuang [3], Yi Yang [2], Qiang Zhang[4], Huajun Chen[2,3*]

[1]The Polytechnic Institute, Zhejiang University
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[3]College of Computer Science and Technology, Zhejiang University
[4]ZJU-UIUC Institute, Zhejiang University
{22460360, dingkeyan, wuyihang, zhuangxiang, y_yi, qiang.zhang.cs, huajunsir}@zju.edu.cn

Abstract

Retrieving molecular structures from tandem mass spectra is a crucial step in rapid compound identification. Existing retrieval methods, such as traditional mass spectral library matching, suffer from limited spectral library coverage, while recent cross-modal representation learning frameworks often encounter modality misalignment, resulting in suboptimal retrieval accuracy and generalization. To address these limitations, we propose GLMR, a Generative Language Model-based Retrieval framework that mitigates the cross-modal misalignment through a two-stage process. In the pre-retrieval stage, a contrastive learning-based model identifies top candidate molecules as contextual priors for the input mass spectrum. In the generative retrieval stage, these candidate molecules are integrated with the input mass spectrum to guide a generative model in producing refined molecular structures, which are then used to re-rank the candidates based on molecular similarity. Experiments on both MassSpecGym and the proposed MassRET-20k dataset demonstrate that GLMR significantly outperforms existing methods, achieving over 40% improvement in top-1 accuracy and exhibiting strong generalizability.

1 Introduction

Tandem mass spectrometry (MS/MS) is one of the most important analytical tools for molecular structure identification (Qiu et al. 2023). In this technique, target molecules are ionized and undergo a multi-stage fragmentation process, generating a set of fragment-ion signals with specific mass-to-charge ratios (m/z). These signals from the fragment mass spectrum reflect the internal chemical bond cleavage patterns and the functional group distribution characteristics of the molecule. In fields such as metabolomics, natural products discovery, and drug development, the accurate retrieval of molecular structures from MS/MS spectra is a fundamental step toward rapid compound identification (Escher, Stapleton, and Schymanski 2020). Herein, the *MS-to-Molecule Retrieval* refers to the process of identifying the most matching molecular structure in a large molecule database based on the input mass spectrum. This retrieval helps researchers quickly locate target compounds (Prudent et al. 2021), eliminating the need for expensive and time-consuming structural analysis experiments (Kind et al. 2018).

However, inferring molecular structures from MS/MS spectra remains a highly challenging problem (Keifer and Jarrold 2017). Firstly, different molecules can generate highly similar spectra, while the same molecule may yield substantially different spectral profiles under varying experimental conditions (El-Aneed, Cohen, and Banoub 2009). Secondly, real-world spectral data often contain noise, missing peaks, or interfering signals (Wang, Li, and Stoica 2005). Conventional approaches primarily employ spectral library matching (Stein and Scott 1994; Kwok, Venkataraghavan, and McLafferty 1973; Wang et al. 2020; Qin et al. 2021), as shown in Figure 1(a), where the experimental spectrum is compared against the reference spectrum of characterized compounds in the databases such as GNPS (Wang et al. 2016), HMDB (Wishart et al. 2022), MoNA (Vaniya et al. 2019) and MassBank (Horai et al. 2010). Although demonstrating reasonable performance for well-documented compounds, these methods exhibit significant limitations due to the restricted coverage of the spectral library (Griss 2016).

Recent advances (Young et al. 2024; Ji et al. 2024; Li et al. 2024) have demonstrated the effectiveness of deep learning approaches in directly learning the intricate relationships between spectral patterns and molecular structures. The most prominent approaches (Goldman et al. 2023b; Kalia et al. 2025) leverage cross-modal representation learning frameworks, as shown in Figure 1(b), where both mass spectra and molecular structures (typically represented as either SMILES strings or molecular graphs) are encoded into a potentially aligned latent space. This paradigm enables efficient MS-to-molecule prediction and retrieval. Despite these advancements, a key challenge persists: *modality misalignment*. Mass spectra and molecular structures belong to fundamentally different modalities: the former describes physical fragmentation behavior, while the latter represents chemical structure information. The gap between them makes

*Corresponding author
†These authors contributed equally to this work.

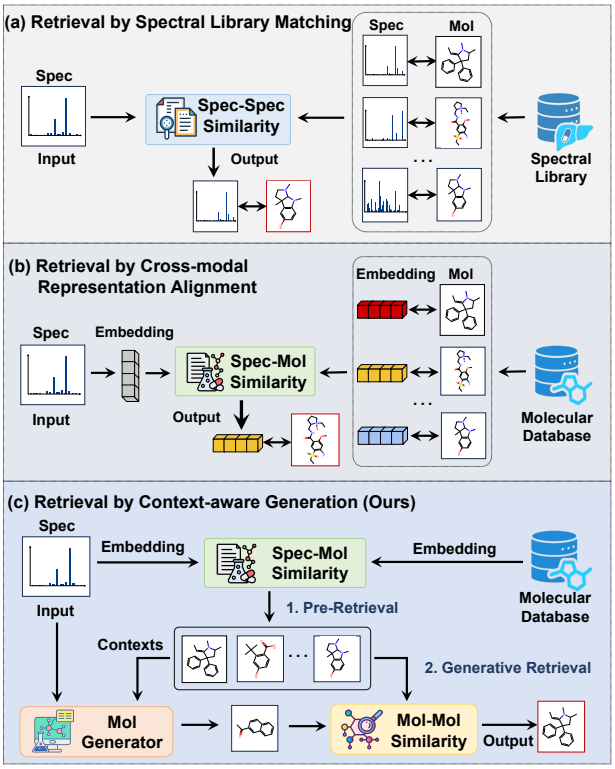

Figure 1: Illustration of the methods for MS-to-Molecule retrieval. (a) **Spectral library matching method**, where the input mass spectrum is compared against the reference MS of characterized compounds in a database. (b) **Cross-modal representation alignment method**, where both mass spectra and molecular structures are encoded into a potentially aligned latent space. (c) **Our method**, which builds upon cross-modal representation alignment, further incorporates a context-aware molecule generator for generative retrieval.

it difficult to establish a well-aligned representation space. As a result, the current state-of-the-art model JESTR (Kalia et al. 2025) demonstrates limited retrieval performance, with top-1 accuracy below 20% in the MassSpecGym benchmark (Bushuiev et al. 2024).

To address this, we propose a generative framework for MS-to-molecule retrieval, termed **GLMR**, as illustrated in Figure 1(c). The core of our approach lies in leveraging a *context-aware generative language model* to bridge the modality gap by generating a molecular structure that is aligned with the input mass spectrum, thereby transforming the cross-modal retrieval into a more tractable unimodal retrieval. Specifically, our method proceeds in two stages: (1) **Pre-Retrieval**: A molecule encoder and a spectral encoder are first trained using contrastive learning to retrieve a set of top-ranked candidate molecules, which serve as contextual priors for the input mass spectrum. (2) **Generative Retrieval**: These candidates, together with spectral features, guide a generative language model to generate a molecular structure aligned with the input mass spectrum. The generated molecule is then compared with the candidate set via molecular similarity, yielding the final retrieval results.

To validate the effectiveness of GLMR, we conduct evaluations not only on MassSpecGym (Bushuiev et al. 2024) but also introduce an enhanced MS-to-molecule retrieval dataset, named **MassRET-20k**, which includes richer spectral variations, providing more challenging and realistic cases. Experimental results on both datasets demonstrate that our method significantly outperforms existing methods. In summary, the main contributions of this study are as follows:

- We propose a novel MS-to-molecule retrieval framework based on generative language models. Our two-stage approach (pre-retrieval and generative retrieval) effectively alleviates the cross-modal misalignment, improving retrieval accuracy and robustness.
- We construct an enhanced molecule retrieval evaluation dataset, enabling comprehensive evaluation of retrieval accuracy, robustness, and generalization.
- Our method achieves over 40% improvement in top-1 accuracy over baseline methods, advancing this field by bridging the gap between mass spectra and molecular structures through generative modeling, enabling more accurate and spectral library-free compound identification.

2 Related Works

2.1 MS-to-Molecule Retrieval Methods

Conventional MS-to-molecule retrieval methods primarily employ spectral library matching, where the input mass spectrum is compared against the reference mass spectrum of known compounds. MASST (Wang et al. 2020) comprises a web-based system for searching the public data repository within the GNPS/MassIVE knowledge base and an analysis infrastructure for a single mass spectrum. DLEAMSE (Qin et al. 2021) introduces a bioinformatics tool enabling rapid spectral retrieval across public repositories and spectral libraries. However, current approaches are fundamentally constrained by the limited availability of spectrum-molecule pairs, with retrieval performance bounded by spectral library coverage. Modeling direct mappings between mass spectra and molecular structures through cross-modal representation learning has emerged as a promising alternative. Contrastive learning has become a prevalent strategy for achieving cross-modal alignment (Khosla et al. 2020). For example, MIST (Goldman et al. 2023a) generates molecular fingerprints based on inferred chemical formulas and performs retrieval via vector similarity, while CMSSP (Chen et al. 2024) integrates molecular graph and fingerprint representations, mapping both spectral and structural modalities into a shared latent space. JESTR (Kalia et al. 2025) further enhances contrastive learning with a candidate molecule regularization strategy, and CSU-MS2 (Xie et al. 2025) improves spectral encoding through sinusoidal m/z embeddings and an enhanced attention module, increasing model expressiveness. However, these methods often suffer from modality misalignment, resulting in suboptimal retrieval accuracy.

2.2 MS-to-Molecule Retrieval Datasets

Spectral libraries such as GNPS (Wang et al. 2016), MoNA (Vaniya et al. 2019), MassBank (Horai et al. 2010), and NIST (Lemmon et al. 2010) serve as foundational resources for MS-to-Molecule retrieval by providing experimentally acquired spectra paired with known molecular structures. However, these datasets often suffer from spectral noise, incomplete metadata, or licensing restrictions, which limit their utility for training and evaluating machine learning models. Several standardized benchmarks such as MIST CANOPUS (Goldman et al. 2023b) and CASMI (Schymanski and Neumann 2013) have been proposed, but they are constrained by small size, potential data leakage, or high preprocessing complexity. Recently, MassSpecGym (Bushuiev et al. 2024) introduced a large-scale, cleaned, and normalized dataset comprising approximately 230k mass spectra, with structurally diverse train-validation-test splits based on MCES (Curchoe 2020) similarity, enabling more robust and reproducible evaluation of retrieval methods.

3 Methodology

The MS-to-molecule retrieval task aims to rank candidate molecules (from a chemical molecule database) based on a given mass spectrum. Formally, given an MS/MS spectrum, the goal is to order a set of candidate molecules such that the correct molecule is positioned at the top (Bushuiev et al. 2024). We address this task through a two-stage retrieval framework (Pre-retrieval and Generative retrieval), as illustrated in Figure 2.

3.1 Pre-retrieval via Cross-modal Representation Alignment

The first stage performs pre-retrieval by aligning molecular and spectral representations to pick a set of candidate molecules. Specifically, we train a cross-modal alignment model via contrastive learning. Following the prior work (Liu et al. 2023), we adopt ChemFormer (Irwin et al. 2022) as the molecular encoder $f_m(\text{Mol};\gamma)$, which is a BART (Lewis et al. 2019) variant pre-trained on the large-scale ZINC database containing billions of compounds (Irwin and Shoichet 2005). Each input molecule is represented as an SMILES sequence, from which the encoder produces a sequence of token embeddings:

$$\mathbf{H}^m = f_m(\text{Mol};\gamma) \in \mathbb{R}^d. \tag{1}$$

A [CLS] token is prepended to the input sequence, and its final hidden state serves as the global molecular embedding $\mathbf{E}^m = \mathbf{H}^m_{\text{[CLS]}}$.

For the spectral encoder $f_s(\text{spectrum};\eta)$, we employ a Transformer architecture with multi-head attention (Vaswani et al. 2017), which allows the model to capture complex relationships across different m/z and intensity dimensions. In contrast to the binning strategies used in previous studies (Kalia et al. 2025; Chen et al. 2024), we represent each mass spectrum as a sequence of tuples $(\text{m/z}, \text{intensity})$, where intensity values are normalized to the range $(0, 1]$. This sequence is then encoded into a set of hidden representations:

$$\mathbf{H}^s = f_s(\text{Spec};\eta) \in \mathbb{R}^d. \tag{2}$$

We apply average pooling over the dimension of sequence length to obtain a fixed-size representation of the mass spectrum, i.e., $\mathbf{E}^s = \frac{1}{T}\mathbf{H}^s$, with T denoting the number of spectral peaks.

The training objective is to align molecular and spectral representations in a latent space. Following the CLIP framework (Radford et al. 2021), we optimize a dual-path InfoNCE loss that encourages mutual alignment between the two modalities. Given a batch of spectrum-molecule pairs $\mathcal{B}$, for the molecule-to-MS alignment, we consider the pairs as the positive sample $\mathbf{E}^s_i$ and construct N negative samples $\mathbf{E}^s_j$ by applying random intensity perturbations to spectral peaks. This yields the molecule-to-MS loss $\mathcal{L}_{mol2ms}$. For the MS-to-molecule alignment, we use the pairs as the positive sample $\mathbf{E}^m_i$ and sample M negative examples $\mathbf{E}^m_j$ from other molecules in the same batch, resulting in the MS-to-molecule loss $\mathcal{L}_{ms2mol}$. The final training loss $\mathcal{L}_{\text{pre}}$ is computed as the average of these two components, defined as:

$$\mathcal{L}_{mol2ms} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{\exp(\frac{\mathbf{E}^m_i\cdot\mathbf{E}^s_i}{\tau})}{\sum_{j=1}^{N}\exp(\frac{\mathbf{E}^m_i\cdot\mathbf{E}^s_j}{\tau})}, \tag{3}$$

$$\mathcal{L}_{ms2mol} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{\exp(\frac{\mathbf{E}^m_i\cdot\mathbf{E}^s_i}{\tau})}{\sum_{j=1}^{M}\exp(\frac{\mathbf{E}^s_i\cdot\mathbf{E}^m_j}{\tau})}, \tag{4}$$

$$\mathcal{L}_{\text{pre}} = \frac{1}{2}(\mathcal{L}_{ms2mol} + \mathcal{L}_{mol2ms}), \tag{5}$$

where τ is the temperature coefficient that controls the sharpness of the similarity distribution.

After training, we use the learned encoders to retrieve candidate molecules in a database based on the cosine similarity between the spectral embedding and each molecular embedding:

$$\mathbf{R}(\mathbf{E}^s, \mathbf{E}^m_i) = \cos(\mathbf{E}^s, \mathbf{E}^m_i) = \frac{\mathbf{E}^s\cdot\mathbf{E}^m_i}{|\mathbf{E}^s|\cdot|\mathbf{E}^m_i|}. \tag{6}$$

The top-K molecules with the highest similarity scores are selected as the output of the pre-retrieval stage, serving as contextual priors to guide the generation of refined molecules in the next stage.

3.2 Generative Retrieval via Context-aware Molecule Generation

The second stage leverages a generative language model conditioned on both the input mass spectrum and the prior candidate molecules to produce refined molecular structures. These generated structures are then used to re-rank the candidate molecules based on molecular similarity, yielding the top-ranked molecules as final retrieval results.

To maintain architectural consistency with the molecular encoder, we employ the ChemFormer Decoder (Irwin et al. 2022) for molecular generation. The input spectrum is encoded as $\mathbf{H}^s$ using the spectral encoder f_s, while the top-K candidate molecules from the pre-retrieval stage are encoded as $\mathbf{H}^m_K = \{\mathbf{H}^m_i\}_{i=1}^K$ using the molecular encoder f_m,

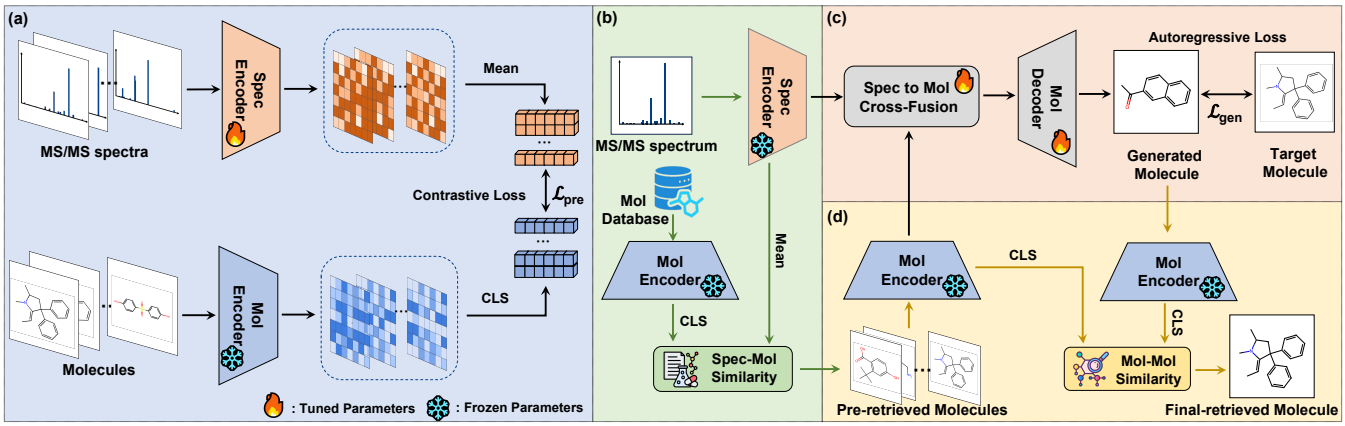

Figure 2: Overview of the proposed GLMR method. (a) **Training process of modality alignment**. We optimize a contrastive loss that encourages mutual alignment between the molecular and spectral modalities. (b) **Inference process of pre-retrieval**. We use the learned encoders to rank candidate molecules in the retrieval database. The top-K molecules with the highest similarity scores are selected as the output of the pre-retrieval stage. (c) **Training process of generative language models**. We leverage a generative language model conditioned on both the input mass spectrum and the prior candidate molecules to produce refined molecular structures. (d) **Inference process of generative retrieval**. We use the generated molecule to re-rank the pre-retrieved molecules based on molecular similarity.

where K denotes the number of pre-retrieved molecules. To effectively integrate the spectral and molecular representations, we introduce a *Cross-Fusion* module, which employs cross-attention mechanisms to fuse $\mathbf{H}^s$ and $\mathbf{H}_K^m$. The resulting fused embedding $\mathbf{H}$ is defined as:

$$\begin{aligned}\mathbf{H}^{\text{ca}} &= f_{\text{ca}}\left(\mathbf{H}^s, \mathbf{H}_K^m; \theta\right) \\ &= \text{Attn}\left(\text{Query}\left(\mathbf{H}^s\right), \text{Key}\left(\mathbf{H}_K^m\right)\right) \cdot \text{Value}\left(\mathbf{H}_K^m\right),\end{aligned} \tag{7}$$

where f_{ca} denotes the cross-attention function parameterized by θ. The Attn function computes cross-attention weights using Query, Key, and Value matrices, which are linear transformations of the input embeddings. This module enables the model to selectively attend to informative molecular candidates while conditioning on the input spectrum.

During training, both the spectral encoder and molecular encoder are kept frozen to preserve the pre-trained cross-modal alignment learned in the pre-retrieval stage. Training focuses solely on the fusion module and the decoder. The generative model $f_g(\mathbf{H}^{\text{ca}}; \phi)$ autoregressively produces the target molecular structure in the form of a SMILES string $y = (y_1, y_2, \ldots, y_Q)$ by maximizing the conditional likelihood of the ground-truth molecule given the fused representation. The training objective is defined as:

$$\mathcal{L}_{\text{gen}} = -\frac{1}{|\mathcal{B}| \cdot Q} \sum_{i=1}^{|\mathcal{B}|} \sum_{q=1}^{Q} \log P\left(y_{i(q)} \mid y_{i(<q)}, \mathbf{H}_i^{\text{ca}}\right), \tag{8}$$

where $P\left(y_{i(q)} \mid y_{i(<q)}, \mathbf{H}_i^{\text{ca}}\right)$ is the conditional probability of token $y_{i(q)}$ at position q, given the preceding tokens $y_{i(<q)}$ and the fused embedding $\mathbf{H}_i^{\text{ca}}$. Q is the target sequence length.

After training, the model is employed to generate a refined molecule, which is then used to re-rank the pre-retrieved molecules based on molecular similarity. Specifically, the generated molecule is encoded as $\mathbf{E}_+^m$, while the candidate molecules are encoded as $\mathbf{E}_i^m$ for $i \in [1, K]$. We calculate cosine similarity between the embeddings of the generated molecule and each candidate:

$$\mathbf{R}\left(\mathbf{E}_+^m, \mathbf{E}_i^m\right) = \cos\left(\mathbf{E}_+^m, \mathbf{E}_i^m\right) = \frac{\mathbf{E}_+^m \cdot \mathbf{E}_i^m}{|\mathbf{E}_+^m| \cdot |\mathbf{E}_i^m|}. \tag{9}$$

The candidate molecules are then re-ranked according to their similarity scores, producing the final output of the generative retrieval stage.

4 Experiments

This section presents a comprehensive evaluation of the retrieval performance of GLMR. We first detail the experimental settings, including datasets, baselines, training configuration, and evaluation criteria. We then report performance on the MassSpecGym and MassRET-20k benchmarks, followed by analyses of modality alignment and generation ability. Lastly, we conduct an ablation study to quantify the contribution of each stage and assess the sensitivity to the number of pre-retrieved candidates.

4.1 Experimental Settings

Datasets We evaluate GLMR on two benchmark datasets: MassSpecGym (Bushuiev et al. 2024) and MassRET-20k. (1) **MassSpecGym** provides two retrieval libraries for each MS/MS spectrum. The first is based on molecular *weight* inferred from the precursor m/z, and the second leverages chemical *formula* matching. (2) To better evaluate model performance under diverse experimental conditions, we construct a new benchmark dataset, **MassRET-20k**. To avoid data leakage, we exclude molecules that appear in the MassSpecGym training set, resulting in a clean evaluation

Library Type	Method	Recall↑ (%)			MRR↑ (%)	MCES@1↓
		Recall@1	Recall@5	Recall@20		
Weight-based Retrieval Library	Random	0.296	1.874	7.684	1.319	31.01
	DeepSets	1.117	4.049	13.459	3.923	25.47
	Fingerprint FFN	3.076	9.211	22.699	7.477	23.85
	DeepSets + Fourier features	9.028	21.081	38.898	15.679	20.87
	MIST	18.455	40.009	64.388	29.302	<u>15.37</u>
	JESTR	<u>17.617</u>	<u>40.355</u>	<u>64.764</u>	<u>29.121</u>	15.82
	GLMR (Ours)	**64.172**	**72.961**	**78.782**	**67.817**	**11.14**
Formula-based Retrieval Library	Random	2.470	10.584	21.251	5.411	13.51
	DeepSets	4.699	12.355	29.289	9.901	13.12
	Fingerprint FFN	4.978	15.505	33.168	11.193	13.09
	DeepSets + Fourier features	10.104	22.015	40.681	16.967	13.01
	MIST	10.942	23.815	44.634	18.257	12.75
	JESTR	<u>11.772</u>	<u>33.258</u>	<u>61.006</u>	<u>22.825</u>	<u>11.73</u>
	GLMR (Ours)	**68.478**	**78.087**	**84.216**	**72.472**	**5.05**

Table 1: Retrieval Performance on **MassSpecGym**. The best results are in bold, and the results ranked second are underlined.

set of approximately 20k spectrum-molecule pairs. Compared to the metadata of MassSpecGym, which includes only two ionization adducts, and where only 53% of the data provides normalized collision energy, resulting in incomplete data, our constructed dataset includes 12 ionization adducts, where all entries include normalized collision energy. As a result, our dataset provides more comprehensive information and more accurately reflects real-world scenarios. Furthermore, the same molecule exhibits different mass spectra under different ionization adducts, making MassRET-20k more challenging than MassSpecGym.

Training Setup Our model is trained in two stages using spectrum-molecule pairs from the MassSpecGym training set. In the pre-retrieval stage, we initialize the molecular encoder with pre-trained ChemFormer weights and randomly initialize the spectral encoder. The model is trained for 300 epochs using contrastive loss, with only the spectral encoder updated while the molecular encoder is frozen. In the generative retrieval stage, we further initialize the decoder with pre-trained weights from ChemFormer and randomly initialize the cross-fusion module. Training runs for 30 epochs with the encoders frozen, updating only the fusion module and decoder to generate molecules conditioned on spectral and contextual information.

Baselines We compare GLMR against a range of baselines spanning traditional and deep learning approaches for MS-to-molecule retrieval. These include Fingerprint FFN (Rumelhart et al. 1986), DeepSets (Zaheer et al. 2017), DeepSets with Fourier features (Zaheer et al. 2017), MIST (Goldman et al. 2023b), and JESTR (Kalia et al. 2025), the current state-of-the-art method for mass spectrum-based cross-modal molecular retrieval.

Evaluation Metrics We employ three standard metrics to evaluate MS-to-molecule retrieval performance: (i) **Recall@K** measures the proportion of test samples for which the ground-truth molecule appears within the top-K ranked candidates. We report Recall@1, Recall@5, and Recall@20 as percentages. (ii) **MRR** (Mean Reciprocal Rank) captures the average inverse rank of the first correct match, giving higher weight to models that rank the true molecule more highly. (iii) **MCES@1** (Maximum Common Edge Subgraph similarity at rank 1) evaluates structural similarity between the top-1 predicted molecule and the ground truth.

4.2 Main Results

Performance on MassSpecGym Table 1 presents the retrieval performance on the MassSpecGym benchmark. GLMR consistently outperforms all baseline methods across all metrics. On the weight-based and formula-based retrieval tasks, GLMR achieves a remarkable improvement in Recall@1, 46% and 56% respectively, over the previous state-of-the-art method JESTR. The significant reduction in MCES@1 further demonstrates that the top-1 predictions from GLMR are structurally closer to the ground truth, even when the exact match is not retrieved. This performance leap stems from GLMR's two-stage design, which decouples retrieval into pre-retrieval (cross-modal alignment) and generative retrieval (context-aware generation). While prior methods encode molecules and mass spectra into a potentially aligned latent space (similar to the pre-retrieval in GLMR), their cross-modal alignment capability remains limited. GLMR addresses this by generating a molecule conditioned on the input mass spectrum and pre-retrieved candidates, thereby reframing the task as a unimodal retrieval process that effectively bridges the modality gap.

Performance on MassRET-20k Table 2 shows the generalization ability of all models on the proposed MassRET-20k benchmark, a more challenging and realistic dataset with diverse ionization adducts and complete experimental metadata. All models are trained solely on the MassSpecGym training set and evaluated in a zero-shot (Pourpanah et al. 2022) transfer setting, making this a rigorous test of *generalization*. As shown in Table 2, GLMR remains the top-

Library Type	Method	Recall↑ (%)			MRR↑ (%)	MCES@1↓
		Recall@1	Recall@5	Recall@20		
Weight-based Retrieval Library	Random	0.330	1.864	7.605	1.329	30.61
	DeepSets	0.986	3.707	12.540	3.902	26.40
	Fingerprint FFN	3.733	12.379	25.705	8.898	22.48
	DeepSets + Fourier features	8.002	20.467	38.140	14.962	21.69
	MIST	14.388	36.417	63.526	25.499	18.71
	JESTR	16.490	38.450	60.636	27.454	18.03
	GLMR (Ours)	**54.042**	**64.347**	**72.984**	**58.835**	**12.08**
Formula-based Retrieval Library	Random	2.012	9.094	26.139	5.956	13.51
	DeepSets	2.729	9.116	27.564	6.737	13.39
	Fingerprint FFN	2.912	10.573	28.322	8.201	13.15
	DeepSets + Fourier features	4.621	14.063	34.497	10.927	13.11
	MIST	6.779	16.882	34.290	12.780	13.04
	JESTR	7.440	23.314	47.356	16.282	11.86
	GLMR (Ours)	**51.141**	**60.062**	**70.671**	**55.565**	**6.94**

Table 2: Retrieval Performance on MassRET-20k. The best results are in bold, and the results ranked second are underlined.

performing method, significantly outperforming all baselines. Compared to baselines, GLMR shows superior generalization on both retrieval libraries, with notably higher Recall@K and MRR, as well as lower MCES@1 scores. These gains underscore GLMR's ability to generalize to unseen mass spectra and varying ionization conditions.

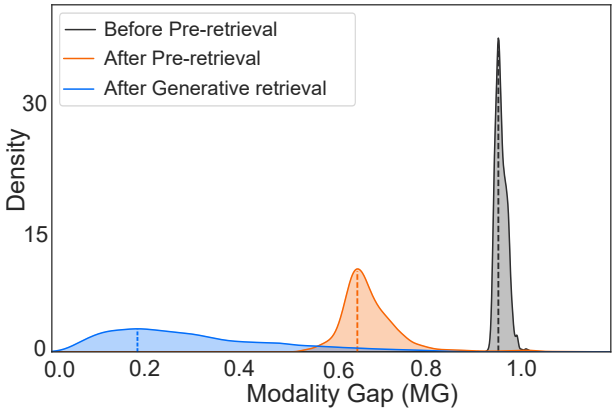

Figure 3: The modality gap distributions on MassSpecGym. A smaller MG indicates better alignment between MS/MS spectra and molecules.

Analysis of Modality Alignment To evaluate whether our method successfully improves modality alignment between MS/MS spectra and molecules, we define a modality gap metric for each instance i as $\text{MG}(\mathbf{E}_i) = 1 - \cos(\mathbf{E}_i^t, \mathbf{E}_i^m)$, where $\mathbf{E}_i^m$ is the representation of the ground-truth molecule, and $\mathbf{E}_i^t$ is the representation of the input mass spectrum (before or after pre-retrieval) or the generated molecule (after generative retrieval). Figure 3 shows the kernel density estimation (Parzen 1962) of the modality gap distribution on MassSpecGym. After the pre-retrieval stage, the distribution shifts leftward compared to the initial distribution, indicating improved alignment through contrastive learning. More notably, in the generative retrieval stage, the modality gap is further and significantly reduced, as the generated molecule is structurally and semantically refined to align closely with the generated molecule. This progressive narrowing of the modality gap demonstrates that GLMR effectively bridges the gap between mass spectra and molecular structures, validating the core advantage of our two-stage framework.

Method	MCES ↓	Morgan Tanimoto ↑	RDK Tanimoto ↑
SMILES-Trans	79.39	0.07	0.03
SELFIES-Trans	33.28	0.10	0.08
SPEC2MOL	37.76	0.12	0.19
$\text{MADGEN}_{\text{Pred}}$	74.19	0.08	0.13
DiffMS	**18.45**	**0.28**	**0.49**
Ours	21.83	0.21	0.42

Table 3: Molecular generation performance on the test set of MassSpecGym.

Analysis of Molecular Generation While the primary goal of GLMR is accurate MS-to-molecule retrieval, the quality of generated molecules, is critical to the final retrieval performance. To evaluate the generation capability of our generative model, we employ the MassSpecGym test set and calculate the structural similarity between generated and ground-truth molecules using three metrics: MCES (Curchoe 2020), Morgan Tanimoto (Vogt and Bajorath 2020), and RDK Tanimoto (Scalfani, Patel, and Fernandez 2022). We compare our generative model against several methods for de-novo molecule generation from MS/MS spectra, including SMILES Transformer (Sennrich, Haddow, and Birch 2015), SELFIES Transformer (Krenn et al. 2020), Spec2Mol (Litsa et al. 2021), $\text{MADGEN}_{\text{Pred}}$ (Wang

Library Type	Method	Recall↑ (%)			MRR↑ (%)	MCES@1↓
		Recall@1	Recall@5	Recall@20		
Weight-based Retrieval Library	w/o Generative retrieval	20.341	52.789	74.630	32.190	22.45
	w/o Pre-retrieval	41.501	59.313	73.279	49.714	18.92
	GLMR (Ours)	**64.172**	**72.961**	**78.782**	**67.817**	**11.14**
Formula-based Retrieval Library	w/o Generative retrieval	46.030	67.925	83.202	55.900	7.83
	w/o Pre-retrieval	52.968	70.460	83.214	60.805	7.27
	GLMR (Ours)	**68.478**	**78.087**	**84.216**	**72.472**	**5.05**

Table 4: Ablation study on the two-stage retrieval strategy of GLMR on MassSpecGym. The best results are in bold.

et al. 2025), and DiffMS (Bohde et al. 2025). Their results are taken directly from the original reports of DiffMS and MADGEN, and summarized in Table 3. Our generative model achieves competitive performance, ranking second only to DiffMS (SOTA). We attribute this strong generation quality to two key design choices: (1) a pre-trained spectral encoder derived from the cross-modal contrastive learning, and (2) a context-aware generation framework that conditions the decoder on top-ranked candidate molecules from the pre-retrieval stage. As a result, the generated molecules are both spectrally consistent and chemically plausible, which in turn enhances the accuracy of the final ranking.

4.3 Ablation Study

In this section, we conduct ablation studies to analyze the contribution of key components in GLMR and to evaluate the sensitivity of performance to the number of retrieved candidates K in the pre-retrieval stage. Specifically, we first investigate how the pre-retrieval and generative retrieval stages individually contribute to overall retrieval effectiveness. The results are reported in Table 4. One can observe that both the pre-retrieval and generative retrieval stages contribute significantly to the performance of GLMR. When used independently, generative retrieval outperforms pre-retrieval alone, indicating that generating a refined molecule is more effective than direct cross-modal matching. However, the best performance is achieved when both stages are combined, demonstrating their complementary nature: the pre-retrieval provides high-quality molecule priors, while the generative retrieval refines these candidates through explicit molecule generation, leading to significantly improved ranking accuracy.

We further investigate the impact of the number of pre-retrieved molecules (K) on retrieval performance. The results (Figure 4) reveal that most metrics reach a plateau when $K > 40$. While increasing K beyond this threshold yields marginal improvements in retrieval accuracy, it also introduces higher computational costs. Based on this analysis, we select $K = 40$ as the optimal number of pre-retrieved molecules for the pre-retrieval stage, striking a balance between performance gains and computational efficiency.

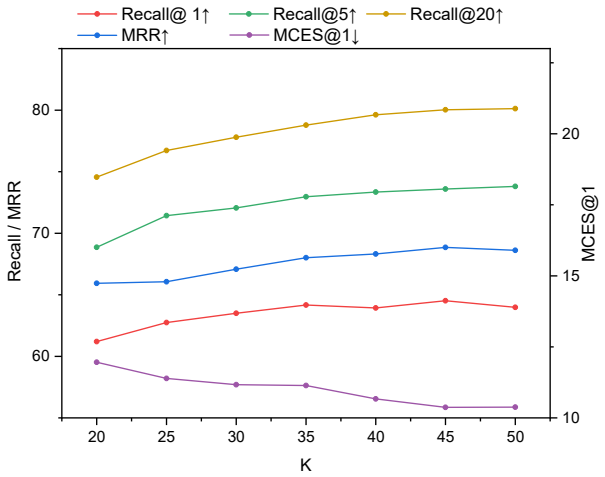

Figure 4: Performance trends with varying K. Experiments are conducted on the weight-based retrieval library of MassSpecGym.

5 Conclusion

In this work, we present GLMR, a generative language model-based framework for MS-to-molecule retrieval that addresses the fundamental challenge of cross-modal misalignment between MS/MS spectra and molecular structures. Our two-stage approach, pre-retrieval and generative retrieval, effectively bridges the modality gap by transforming the inherently challenging cross-modal retrieval into a more tractable unimodal molecule retrieval process. Extensive experiments have demonstrated that our method significantly outperforms existing baselines and shows strong generalization. Despite these advances, GLMR has certain limitations. First, the two-stage pipeline incurs higher computational cost compared to end-to-end retrieval models. Second, while the generative decoder produces structurally plausible molecules, its performance still depends on the quality of the initial candidates. Looking forward, we envision several directions for improvement: designing lightweight fusion and generation modules for faster inference; and incorporating explicit chemical constraints or syntactic rules during generation to enhance molecular structure validity. By combining generative modeling with retrieval, GLMR opens a promising pathway toward accurate, robust, and library-free compound identification in real-world mass spectrometry applications.

References

Bohde, M.; Manjrekar, M.; Wang, R.; Ji, S.; and Coley, C. W. 2025. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. *arXiv preprint arXiv:2502.09571*.

Bushuiev, R.; Bushuiev, A.; de Jonge, N.; Young, A.; Kretschmer, F.; Samusevich, R.; Heirman, J.; Wang, F.; Zhang, L.; Dührkop, K.; et al. 2024. MassSpecGym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37: 110010–110027.

Chen, L.; Xia, B.; Wang, Y.; Huang, X.; Gu, Y.; Wu, W.; and Zhou, Y. 2024. CMSSP: A Contrastive Mass Spectra-Structure Pretraining Model for Metabolite Identification. *Analytical Chemistry*, 96(42): 16871–16881.

Curchoe, C. L. 2020. All models are wrong, but some are useful. *Journal of Assisted Reproduction and Genetics*, 37: 2389–2391.

El-Aneed, A.; Cohen, A.; and Banoub, J. 2009. Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. *Applied spectroscopy reviews*, 44(3): 210–230.

Escher, B. I.; Stapleton, H. M.; and Schymanski, E. L. 2020. Tracking complex mixtures of chemicals in our changing environment. *Science*, 367(6476): 388–392.

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

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

Griss, J. 2016. Spectral library searching in proteomics. *Proteomics*, 16(5): 729–740.

Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7): 703–714.

Irwin, J. J.; and Shoichet, B. K. 2005. ZINC- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1): 177–182.

Irwin, R.; Dimitriadis, S.; He, J.; and Bjerrum, E. J. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022.

Ji, H.; Du, R.; Dai, Q.; Su, M.; Lyu, Y.; Peng, Y.; and Yan, J. 2024. DeepMASS: Unknown Compound Annotation using Semantic Similarity of Mass Spectral Language and Chemical Space Localization. *bioRxiv*, 2024–05.

Kalia, A.; Zhou Chen, Y.; Krishnan, D.; and Hassoun, S. 2025. JESTR: J oint E mbedding S pace T echnique for R anking Candidate Molecules for the Annotation of Untargeted Metabolomics Data. *Bioinformatics*, btaf354.

Keifer, D. Z.; and Jarrold, M. F. 2017. Single-molecule mass spectrometry. *Mass spectrometry reviews*, 36(6): 715–733.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; et al. 2018. Identification of small molecules using accurate mass MS/MS search. *Mass spectrometry reviews*, 37(4): 513–532.

Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; and Aspuru-Guzik, A. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4): 045024.

Kwok, K.-S.; Venkataraghavan, R.; and McLafferty, F. 1973. Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system. *Journal of the American Chemical Society*, 95(13): 4185–4194.

Lemmon, E. W.; Huber, M. L.; McLinden, M. O.; et al. 2010. NIST standard reference database 23. *Reference fluid thermodynamic and transport properties (REFPROP), version*, 9.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X.; Zhou Chen, Y.; Kalia, A.; Zhu, H.; Liu, L.-p.; and Hassoun, S. 2024. An Ensemble Spectral Prediction (ESP) model for metabolite annotation. *Bioinformatics*, 40(8): btae490.

Litsa, E.; Chenthamarakshan, V.; Das, P.; and Kavraki, L. 2021. Spec2Mol: An end-to-end deep learning framework for translating MS/MS Spectra to de-novo molecules.

Liu, S.; Nie, W.; Wang, C.; Lu, J.; Qiao, Z.; Liu, L.; Tang, J.; Xiao, C.; and Anandkumar, A. 2023. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12): 1447–1457.

Parzen, E. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3): 1065–1076.

Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; Wang, X.-Z.; and Wu, Q. J. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4051–4070.

Prudent, R.; Annis, D. A.; Dandliker, P. J.; Ortholand, J.-Y.; and Roche, D. 2021. Exploring new targets and chemical space with affinity selection-mass spectrometry. *Nature Reviews Chemistry*, 5(1): 62–71.

Qin, C.; Luo, X.; Deng, C.; Shu, K.; Zhu, W.; Griss, J.; Hermjakob, H.; Bai, M.; and Perez-Riverol, Y. 2021. Deep learning embedder method and tool for mass spectra similarity search. *Journal of proteomics*, 232: 104070.

Qiu, S.; Cai, Y.; Yao, H.; Lin, C.; Xie, Y.; Tang, S.; and Zhang, A. 2023. Small molecule metabolites: discovery of

biomarkers and therapeutic targets. *Signal Transduction and Targeted Therapy*, 8(1): 132.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Rumelhart, D. E.; McClelland, J. L.; Group, P. R.; et al. 1986. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Scalfani, V. F.; Patel, V. D.; and Fernandez, A. M. 2022. Visualizing chemical space networks with RDKit and NetworkX. *Journal of Cheminformatics*, 14(1): 87.

Schymanski, E. L.; and Neumann, S. 2013. CASMI: and the winner is.... *Metabolites*, 3(2): 412–439.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Stein, S. E.; and Scott, D. R. 1994. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9): 859–866.

Vaniya, A.; Mehta, S.; Wohlgemuth, G.; and Fiehn, O. 2019. MassBank of North America: using untargeted metabolomics and multistage fragmentation mass spectral libraries to annotate natural products in plants. *Berichte aus dem Julius Kühn-Institut*, (204).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vogt, M.; and Bajorath, J. 2020. ccbmlib–a Python package for modeling Tanimoto similarity value distributions. *F1000Research*, 9: Chem–Inf.

Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature biotechnology*, 34(8): 828–837.

Wang, M.; Jarmusch, A. K.; Vargas, F.; Aksenov, A. A.; Gauglitz, J. M.; Weldon, K.; Petras, D.; da Silva, R.; Quinn, R.; Melnik, A. V.; et al. 2020. Mass spectrometry searches using MASST. *Nature biotechnology*, 38(1): 23–26.

Wang, Y.; Chen, X.; Liu, L.; and Hassoun, S. 2025. MADGEN–Mass-Spec attends to De Novo Molecular generation. *arXiv preprint arXiv:2501.01950*.

Wang, Y.; Li, J.; and Stoica, P. 2005. *Spectral analysis of signals: The missing data case*. Morgan & Claypool Publishers.

Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; et al. 2022. HMDB 5.0: the human metabolome database for 2022. *Nucleic acids research*, 50(D1): D622–D631.

Xie, T.; Zhang, H.; Yang, Q.; Sun, J.; Wang, Y.; Long, J.; Zhang, Z.; and Lu, H. 2025. CSU-MS2: A Contrastive Learning Framework for Cross-Modal Compound Identification from MS/MS Spectra to Molecular Structures. *Analytical Chemistry*.

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

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; and Smola, A. J. 2017. Deep sets. *Advances in neural information processing systems*, 30.