MSAnchor: *De Novo* Molecular Generation from Mass Spectrometry Data with Anchor-Extended Molecular Scaffolds

Xiaohan Qin[1,2], Chao Wang[3], Zhengyang Zhou[1,2], Linjiang Chen[1,2], Wenjie Du[1,2*], Yang Wang[1,2*]

[1]University of Science and Technology of China (USTC), Hefei, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3]ByteDance, Inc.
qxh@mail.ustc.edu.cn, wangchao.hhhhh@bytedance.com, {linjiangchen, zzy0929, duwenjie, angyan}@ustc.edu.cn

Abstract

Tandem mass spectrometry (MS/MS) is a critical tool for identifying molecular structures. By efficiently separating molecular fragments based on their mass-to-charge (m/z) ratios, it facilitates molecular generation and subsequent scientific discoveries. However, *de novo* molecular generation from MS/MS spectra remains fundamentally constrained by two paramount challenges: the vast chemical space requires effective structural constraints, and the absence of fine-grained substructural generation weakens the correspondences between spectral features and molecular structures. In this work, we propose **MSAnchor**, a novel two-stage framework for MS/MS-based molecular structure generation. We mitigate the search space challenge through the introduction of Anchor-Extended Molecular Scaffold (AEMS) representation that explicitly encodes side-chain anchoring points, thereby dramatically reducing combinatorial complexity. Leveraging the explicit attachment sites provided by AEMS, we develop anchor-specific priors that establish effective alignments between spectral features and molecular substructures. This fine-grained substructural correspondence is further enhanced by a modified Conditional Information Bottleneck (CIB) module that extracts the most informative spectral components in a structure-aware manner. These innovations enable MSAnchor to generate molecular structures that closely reflect spectral characteristics while constraining combinatorial complexity. Extensive experiments on the CANOPUS and MassSpecGym datasets demonstrate that MSAnchor achieves state-of-the-art performance in molecular structure prediction from MS/MS spectra, with performance improvements that are particularly more pronounced for molecules with higher complexity.

Introduction

Tandem mass spectrometry (MS/MS) is a widely used analytical method for identifying unknown molecules by generating informative fragmentation spectra through precursor ion dissociation (Sleno and Volmer 2004; McLafferty 1981; de Hoffmann 1996). By analyzing the mass-to-charge (m/z) ratios and relative intensities of fragment ions (Winger et al. 1993; McLuckey and Stephenson 1998), MS/MS provides insight into molecular composition and substructures (Aksenov et al. 2017; De Vijlder et al. 2018), thereby supporting compound identification across domains such as metabolomics, drug discovery, and environmental analysis (Beniddir et al. 2021; Wishart 2016). In particular,the precursor ion enables accurate estimation of the molecular formula, providing a strong constraint for molecular generation. (Bristow 2006; Vestal 2001; Luo et al. 2024).

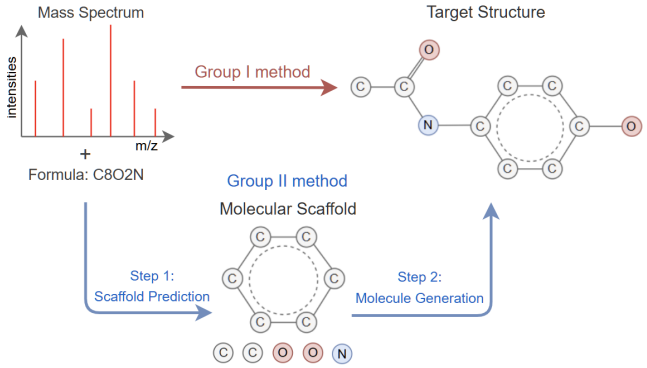

Figure 1: Two paradigms for structure generation from mass spectra: (1) direct mapping to substructures, and (2) scaffold-based generation with reduced complexity.

To reconstruct molecular structures from MS/MS spectra, traditional approaches such as untargeted analysis and database searching have been widely used (Kind et al. 2018; Scheubert, Hufsky, and Böcker 2013; Blaženović et al. 2018), but these methods are often computationally intensive and time-consuming (Nguyen, Nguyen, and Mamitsuka 2019; Vaniya and Fiehn 2015). Recently, AI approaches have enabled the direct mapping from spectral data to molecular structures (Houhou and Bocklitz 2021; Xue et al. 2023). As illustrated in Figure 1, these approaches can generally be categorized into two main groups (Hong, Ye, and Tang 2025; Xue et al. 2023). Group I is an end-to-end paradigm such as neural language models that translate MS/MS directly into SMILES strings. In contrast, Group II comprises two-stage pipelines that first generates intermediate representations, such as scaffolds or molecular fingerprints, and then performs molecular completion (Wang, Li, and Barati Farimani 2023; Zhang et al. 2018; Hu et al. 2023).

*Yang Wang and Wenjie Du are corresponding authors.

Although Group II improves the interpretability of the generation process, it still faces two fundamental challenges. First, due to **the vastness of chemical space** (Coley 2021; Ross et al. 2024), constraints that are limited to molecular scaffolds or fingerprints are often insufficient to effectively narrow the generative space. As a result, models frequently generate molecules that are either chemically invalid or inconsistent with MS/MS spectra (Ma, Vitek, and Nesvizhskii 2012; De Vijlder et al. 2018). Second, while existing models can capture the global coherence of molecular structures (Dean and Scholes 2017; Oprea and Gottfries 2001), they often overlook the **fine-grained generation of molecular substructures**, thereby weakening the structured correspondences between spectral features and specific molecular fragments during generation (Xia et al. 2023; Liyaqat, Ahmad, and Saxena 2024). Moreover, the ambiguity from complex ionization and fragmentation (Niessen et al. 2017) complicates feature extraction and alignment with molecular substructures (Walters and Barzilay 2020).

Here, we propose **MSAnchor**, a novel two-stage framework for *de novo* molecular generation conditioned on **MS**/MS spectra, built upon an extended scaffold representation that we term the **Anchor**-Extended Molecular Scaffold (AEMS). MSAnchor first leverages a Transformer-based model to predict AEMS, then employs a diffusion model for comprehensive molecular structure generation. To reduce the combinatorial search space, AEMS extends standard SMILES scaffolds by introducing virtual atoms (denoted as $\star$) at potential attachment sites. This design facilitates the identification of side-chain linkage and guides structure reconstruction, thereby substantially reducing the complexity of molecular generation.

Moreover, to enable fine-grained generation of molecular substructures, we approach this challenge from two perspectives. On one hand, the explicit attachment sites provided by AEMS allow MSAnchor to construct anchor-specific priors, thereby facilitating effective alignment between spectral features and molecular substructures. On the other hand, to extract the informative spectral components for this alignment, we employ a modified Conditional Information Bottleneck (CIB) module to compress spectral data in a structure-aware manner(Zhang et al. 2025). This alignment enhances both the interpretability and reliability of molecular generation.

Our contributions can be summarized as follows:

- We present MSAnchor, a novel two-stage framework for *de novo* molecular generation built upon AEMS. We demonstrate that AEMS can effectively constrain the combinatorial search space and efficiently guide diffusion models for molecular structure completion.
- Our method leverages AEMS's explicit attachment sites to construct anchor-specific priors, strengthening the alignment between spectral characteristics and molecular substructures. Furthermore, we demonstrate that the CIB module can effectively extract informative features from mass spectra, thereby improving alignment accuracy.
- On established benchmarks for *de novo* structure generation, MSAnchor achieves superior performance compared to all baseline methods, demonstrating improved accuracy and structural similarity. This validates the practical effectiveness of our approach in computational mass spectrometry and drug discovery applications.

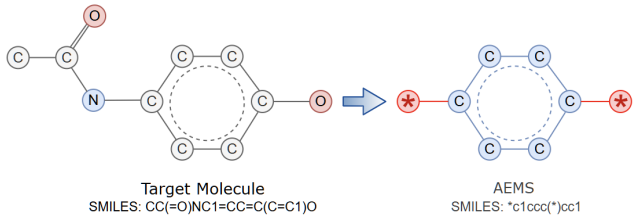

Figure 2: An example of AEMS. The left side shows the target molecule and its SMILES, and the right side shows the corresponding AEMS.

Methodology

In this section, we introduce our method, MSAnchor. We first formally define the AEMS, and then present the model architecture following its two-stage generation strategy.

Anchor-Extended Molecular Scaffold (AEMS)

Standard molecular scaffolds are usually derived by removing all side chains from a molecule, retaining only the core ring systems and linkers. However, this representation loses information on potential branching sites, limiting its effectiveness in structure generation. This limitation becomes particularly problematic when reconstructing molecules with complex or asymmetric side chains, where the exact attachment locations are crucial for accurate prediction. Here, we propose the **Anchor-Extended Molecular Scaffold (AEMS)**, a scaffold representation that preserves explicit attachment points by linking virtual atoms (denoted as "$\star$") to scaffold atoms where side chains are attached. Importantly, this design preserves the molecular representation format like SMILES, ensuring that AEMS remains a valid and parsable molecular string suitable for downstream processing. Figure 2 presents a sample molecule along with its SMILES representation and the corresponding AEMS, clearly illustrating how side-chain anchoring points are retained in the scaffold.

Stage 1: AEMS Prediction with Transformer

To retrieve AEMS, we propose a sequence-to-sequence Transformer model that takes a molecular formula and its corresponding mass spectrum as input and outputs the AEMS in SMILES format. To construct training examples, the molecular formula is tokenized into atomic and numeric symbols, while the mass spectrum is converted into a sequential representation by selecting the top 100 peaks and encoding each as a pair of m/z ratio and relative intensity. These two modalities are strategically concatenated to form the model input, enabling effective cross-modal learning. To obtain the AEMS representation, we first compute the Murcko scaffold of the given molecule. We then identify side-chain attachment points by aligning the scaffold with the molecular graph, and attach virtual atoms at these positions.

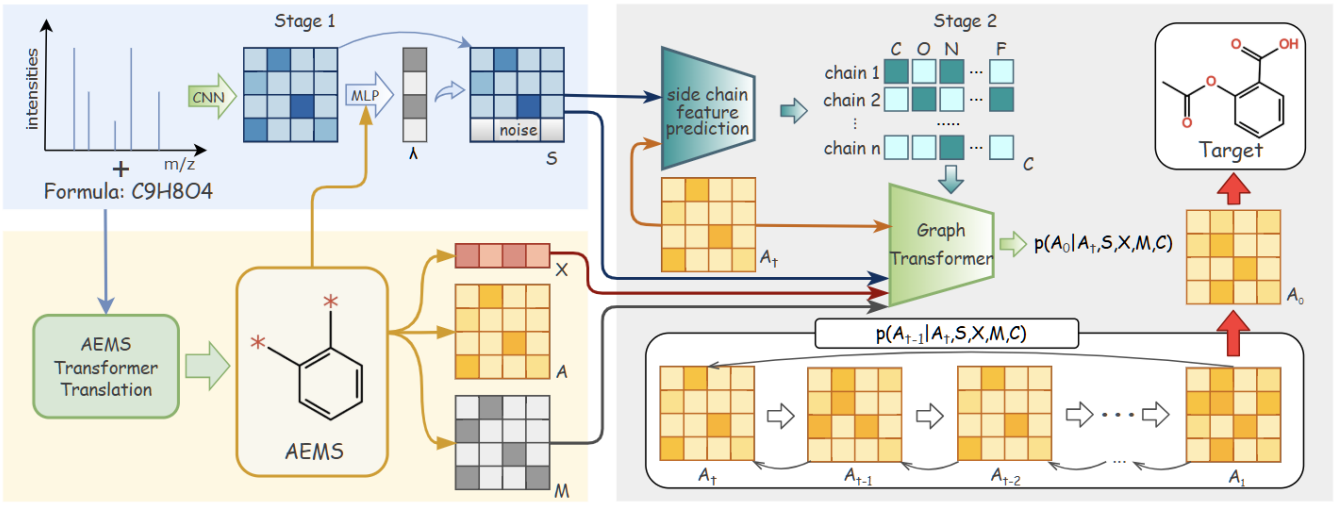

Figure 3: Overview of the MSAnchor framework. Given an input mass spectrum and molecular formula, MSAnchor generates a AEMS scaffold via a Transformer model. The AEMS is then converted into node features X, an adjacency matrix A, and a scaffold mask M. A CIB module compresses the spectrum into a noise-injected representation S. A side-chain predictor outputs atom compositions C for each branch. Finally, all features are combined in a diffusion model to reconstruct the molecular A_0.

To predict AEMS, we employ a Transformer-based sequence-to-sequence model that generates candidate AEMS sequences. We utilize beam search with a width of 5 to produce probability-ranked candidate sequences. After AEMS prediction, a post-processing validation framework is employed to enhance prediction reliability. We introduce a hierarchical evaluation protocol that assesses candidates in descending order: first validating chemical validity using cheminformatics rules, then verifying compliance with atomic composition constraints derived from the molecular formula. Invalid candidates are systematically discarded through our filtering mechanism. When no valid candidates remain, we set the scaffold prediction to empty, providing greater flexibility for downstream molecular generation while maintaining chemical plausibility. This validation approach ensures that our predictions are both chemically valid and aligned with the input formula constraints.

Stage 2: Molecule Completion via Spectra-guided AEMS-conditioned Diffusion

Problem Formulation Given a target mass spectrum S and an AEMS $\mathcal{G}_A = (\mathcal{V}_A, \mathcal{E}_A)$, the objective is to generate a complete molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ such that its predicted spectrum closely approximates the target (Du et al. 2025). The AEMS encodes structural priors via two components: a core scaffold representation, defined as a subgraph of the target molecule ($\mathcal{G}_A \subseteq \mathcal{G}$), and a set of designated anchor points that specify attachment sites for side chains.

To implement this conditional generation, a discrete diffusion framework over molecular graphs is adopted. Initially, a molecule is represented by three tensors:

$$X \in \{0,1\}^{N\times d}, \quad A \in \{0,\dots,4\}^{N\times N}, \quad M \in \{0,1\}^{N\times N}. \tag{1}$$

Here, X denotes the node feature matrix, A is the adjacency matrix representing bond types, and M is the scaffold mask indicating which edges must remain unchanged. The detailed processing procedures are provided in Appendix.

Conditional Information Bottleneck on Spectra To enhance the informativeness of spectral representations during molecular complementary generation, we modify a Conditional Information Bottleneck (CIB) module to compress the spectrum $Z = \{z_j\}_{j=1}^{L}$ conditioned on the scaffold representation h_{scaffold} in a structure-aware manner.

The optimization objective of CIB can be formulated as:

$$\mathcal{L}_{\text{CIB}} = -I(S; Y \mid h_{\text{scaffold}}) + \beta I(Z; S \mid h_{\text{scaffold}}), \tag{2}$$

where Z represents the original spectrum, S denotes the compressed spectral representation, Y is the target molecular structure, and h_{scaffold} is the scaffold representation. The first term encourages S to retain more informative features relevant to molecular structure, while the second term aims to compress irrelevant information as much as possible. To achieve effective compression, the key idea is to enable the model to inject noise into insignificant spectral regions while preserving informative frequency bands.

Specifically, a gating network computes token-level importance scores $p_j \in [0,1]$ via a shared MLP. These scores determine which spectral tokens are retained and which are replaced by Gaussian noise which is computed as:

$$s_j = \lambda_j z_j + (1-\lambda_j)\epsilon_j, \quad \epsilon_j \sim \mathcal{N}(\mu_j, \sigma_j^2), \tag{3}$$

where λ_j is obtained through Gumbel-Softmax (Maddison, Mnih, and Teh 2016):

$$\lambda_i = \sigma\left(\frac{1}{t}\log\left(\frac{p_i}{1-p_i}\right) + \log\left(\frac{u}{1-u}\right)\right), \tag{4}$$

where $u \sim \text{Uniform}(0,1)$, and t is the temperature hyperparameter that is set to 1.0 in this work.

To encourage the compressed spectrum S to be both informative for molecular structure prediction and compact with respect to spectrum Z and scaffold h_{scaffold}, we adopt the CIB objective. Specifically, following its variational formulation, two KL-based regularization terms are incorporated into the loss function:

$$\mathcal{L}_{\text{KL1}} = E_{Z,h}\left[-\frac{1}{2}\log A + \frac{1}{2N}A + \frac{1}{2N}B^2\right], \tag{5}$$

$$\mathcal{L}_{\text{KL2}} = E_h\left[-\frac{1}{2}\log A' + \frac{1}{2N}A' + \frac{1}{2N}(B')^2\right], \tag{6}$$

$$A = \sum_{j=1}^{N}(1-\lambda_j)^2, \qquad B = \sum_{j=1}^{N}\lambda_j \cdot \frac{s_j - \mu_S}{\sigma_S},$$

$$A' = \sum_{j=1}^{N}(1-\lambda_j^{(h)})^2, \quad B' = \sum_{j=1}^{N}\lambda_j^{(h)} \cdot \frac{s_j^{(h)} - \mu_S}{\sigma_S}.$$

Here, λ_j and $\lambda_j^{(h)}$ are importance weights computed from the full input and scaffold-only features, respectively; s_j and $s_j^{(h)}$ are token-level embeddings; μ_S and σ_S are the mean and standard deviation of the variational posterior $q(S)$. Detailed derivations are provided in Appendix.

Side-Chain Composition Prediction

The side-chain composition matrix C serves as an intermediate representation that bridges the spectral input S with AEMS. Instead of generating full substructures directly, atomic compositions are first inferred to impose chemical constraints on subsequent molecule generation process.

Given an AEMS $\mathcal{G}_A$ with k anchor points and mass spectrum S, the prediction of side-chain compositions is formulated as a multi-label classification problem. For each anchor $i \in \{1,\ldots,k\}$, a binary vector $c_i \in \{0,1\}^a$ is predicted to indicate the presence of each atom type. These vectors are aggregated to form the matrix $C \in \{0,1\}^{k\times a}$. To integrate both local structural context and global spectral information, a graph-spectral fusion strategy is adopted. Node embeddings H for the scaffold graph are computed using L layers of graph neural networks as follows:

$$H^{(\ell+1)} = \sigma\left(\text{GNN}^{(\ell)}(H^{(\ell)}, \mathcal{E}_A)\right), \tag{7}$$

where $H^{(0)}$ is the initial node feature matrix, and $\text{GNN}^{(\ell)}$ denotes the ℓ-th graph convolution layer, σ denotes an activation function.

The prediction process follows a structured and modular pipeline. First, for each anchor node in the molecular scaffold, the final node embedding obtained from the graph neural network is concatenated with the global spectral feature vector to create a fused representation that captures both local structural information and global spectral characteristics. Then, this fused vector is passed through a prediction-specific multi-layer perceptron, followed by a sigmoid activation function, to generate the predicted atom-type probabilities for each anchor position, enabling multi-label classification of potential attachment atoms.

The prediction head is optimized using a binary cross-entropy loss, applied over all anchor indices $\mathcal{A} \subseteq \{1,\ldots,|\mathcal{V}_A|\}$:

$$\mathcal{L}_C = -\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\sum_{j=1}^{a}\left[y_{ij}\log(c_{ij}) + (1-y_{ij})\log(1-c_{ij})\right], \tag{8}$$

where $y_{ij} \in \{0,1\}$ indicates whether atom type j is present in the side chain attached to anchor i.

Molecular Complementary Generation

Forward Process At each diffusion timestep $t = 1,\ldots,T$, noise is independently applied to each edge according to transition matrices $Q_1,\ldots,Q_T$, where the element $Q_t^{mn} = q(a^t = n \mid a^{t-1} = m)$ denotes the probability of transitioning from bond type m to n at step t. The forward process is defined as

$$q(A^t \mid A^{t-1}) = A^{t-1}Q_t. \tag{9}$$

By the Markov property, the distribution of A^t conditioned on original adjacency matrix A^0 can be expressed as

$$q(A^t \mid A^0) = A^0\bar{Q}_t, \tag{10}$$

where $\bar{Q}_t = Q_1Q_2\cdots Q_t$ is the composed transition matrix. Since molecular graphs are undirected, noise is added only to the upper triangular part of A, after which the matrix is symmetrized.

For categorical bond types, the transition matrix is parameterized as

$$\bar{Q}_t = \bar{\alpha}_t I + \bar{\beta}_t \mathbf{1}m^\top, \tag{11}$$

where m denotes the empirical distribution of bond types estimated from the training data. The noise coefficients $\bar{\alpha}_t$ and $\bar{\beta}_t = 1-\bar{\alpha}_t$ are scheduled following the cosine schedule:

$$\bar{\alpha}_t = \cos^2\left(\frac{\pi(t/T+\varepsilon)}{2(1+\varepsilon)}\right), \quad \varepsilon = 0.008. \tag{12}$$

Reverse Process The reverse denoising process is implemented through a specialized graph neural network ϕ_θ that integrates structural and spectral information. The core of the model is a graph transformer module with edge-aware attention, where node representations are updated by

$$h_i^{(\ell)} = h_i^{(\ell-1)} + \text{MHA}(h_i^{(\ell-1)}, h^{(\ell-1)}, A), \tag{13}$$

where MHA denotes a multi-head attention mechanism that incorporates edge features A to modulate the attention weights. To effectively capture bond relationships, an edge gating function $g(A_{ij})$ biases the attention scores as

$$\alpha_{ij} \propto \exp(Q_i \cdot K_j^T + g(A_{ij})), \tag{14}$$

enabling the model to prioritize structurally relevant connections while respecting molecular topology. Spectral information S, side-chain composition C, and timestep embeddings t are integrated into a condensed conditioning vector

$$c_t = W_c[z_S; t_{\text{emb}}; C], \tag{15}$$

which is then projected and added to node representations.

Retriever	Top1 Acc.↑	Top1 Sim.↑	Top1 MCES↓	Top10 Acc.↑	Top10 Sim.↑	Top10 MCES↓
CANOPUS						
Spec2Mol	0.00%	0.12	27.82	0.00%	0.16	23.13
MADGEN	2.10%	0.22	20.56	2.39%	0.27	12.69
MIST+ND	2.32%	0.35	12.11	6.11%	0.43	9.91
MIST+MSN	5.40%	0.34	14.52	11.04%	0.44	10.23
DiffMS	8.34%	0.35	11.95	15.44%	0.47	9.23
MSAnchor	**8.51%**$_{(0.002)}$	**0.38**$_{(0.006)}$	**11.12**$_{(0.12)}$	**16.90%**$_{(0.003)}$	**0.49**$_{(0.008)}$	**8.95**$_{(0.09)}$
MassSpecGym						
Rand. Gen.	0.00%	0.08	21.11	0.00%	0.11	18.25
SMILES Trans.	0.00%	0.03	79.39	0.00%	0.10	52.13
SELFIES Trans.	0.00%	0.08	38.88	0.00%	0.13	26.87
Spec2Mol	0.00%	0.12	37.76	0.00%	0.16	29.40
MIST+ND	0.00%	0.14	33.19	0.00%	0.16	31.89
MIST+MSN	0.00%	0.06	45.55	0.00%	0.15	30.13
MADGEN	1.31%	0.20	27.47	1.54%	0.26	16.84
DiffMS	2.30%	0.28	18.45	4.25%	0.39	14.73
MSAnchor	**2.68%**$_{(0.017)}$	**0.32**$_{(0.011)}$	**16.57**$_{(0.28)}$	**4.67%**$_{(0.023)}$	**0.41**$_{(0.012)}$	**14.12**$_{(0.26)}$

Table 1: Performance comparison on CANOPUS and MassSpecGym datasets. Best results are in **bold**, second-best results are underlined, and standard deviations are shown as subscripts.

At each diffusion step, an edge prediction module estimates the denoised adjacency matrix while preserving scaffold constraints. Given node embeddings h_i, timestep embedding c_t, and scaffold mask M, the bond type a_{ij}^0 between atoms i and j is predicted by

$$p_\theta(a_{ij}^0 = b \mid A_t, M) = \mathrm{Softmax}\big(f_{\text{edge}}(h_i, h_j, c_t)\big)\,(1 - M_{ij}) + \delta_{b,a_{ij}^0}\,M_{ij}. \quad (16)$$

where f_{edge} is a bond classifier and δ enforces scaffold edges to remain fixed. Training employs a masked cross-entropy loss focusing on edges unconstrained by the scaffold.

Finally, CIB regularization terms as well as side chain prediction loss are incorporated into the training objective:

$$\mathcal{L} = \mathcal{L}_{pred} + \lambda_1 \mathcal{L}_{\mathrm{KL1}} + \lambda_2 \mathcal{L}_{\mathrm{KL2}} + \lambda_3 \mathcal{L}_c, \quad (17)$$

$$\mathcal{L}_{pred} = E_{t,A_t}[\sum_{i,j}(1 - M_{ij}) \cdot \mathrm{CE}(a_{ij}^0, p_\theta)], \quad (18)$$

where $\mathcal{L}_{\mathrm{KL1}}$ and $\mathcal{L}_{\mathrm{KL2}}$ are KL-based regularization terms, $\mathcal{L}_c$ is the prediction loss of side chain composition, and λ_1, λ_2, λ_3 are hyperparameters. This formulation guarantees reconstructed structures honor scaffold topology, align with informative spectral evidence under the CIB framework. Details of the hyperparameter sensitivity analysis are presented in the Appendix.

Experiment and Analyses

Datasets and Setups

Datasets We evaluate our model on **CANOPUS**, a curated dataset of small molecules with one-to-one spectrum–structure pairs (Dührkop et al. 2015), and **MassSpecGym**, the largest unified MS/MS benchmark with structural annotations (Bushuiev et al. 2024). To capture universal chemical features, the first-stage Transformer was pretrained on a chemically diverse dataset combining spectra from the National Institute of Advanced Science and Technology, SDBS Web[1] and the large-scale dataset from Alberts et al. (Alberts et al. 2024). To avoid data leakage, molecules with an MCES distance <11 from any test molecule were excluded. All experiments were conducted using the official dataset splits. Each experiment is repeated 8 times with different random seeds, and we report the mean and standard deviation of the results.

Baselines We compare MSAnchor with several baseline models for *de novo* molecular generation from spectra. These include Group I approaches such as SMILES Transformer, Spec2Mol (Litsa et al. 2021) and DiffMS (Bohde et al. 2025), as well as Group II pipelines like MSNovelist (Stravs et al. 2022), MIST combines with Neuraldecipher (Goldman et al. 2023a; Le et al. 2020), and MADGEN (Wang et al. 2025).

Metrics We evaluate generation performance using three complementary metrics: Top-k accuracy, which checks whether the ground-truth molecule appears among the top-k predictions; Tanimoto similarity, which measures fingerprint-based structural similarity (Morgan 1965); and Maximum Common Edge Substructure (MCES) (Kretschmer et al. 2023), a graph-based metric computing edit distance on the largest common substructure.

Experimental Setup In the first stage, the transformer is pretrained for 100,000 steps and finetuned for 20,000 steps. In the second stage, MSAnchor is trained using the AdamW optimizer with a batch size of 40 and a learning rate of 1e-4. Models are trained for 100 epochs. Implementation is based on the PyTorch Lightning framework and executed on two

[1] https://sdbs.db.aist.go.jp

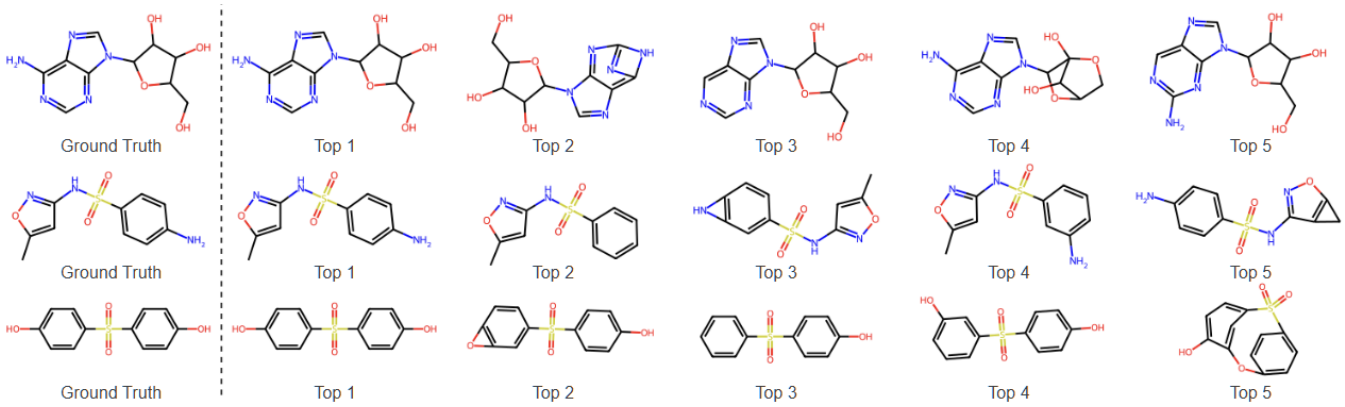

Figure 4: Generation examples by MSAnchor. The left column displays the target molecules, while the right five columns show the model's top-5 predictions.

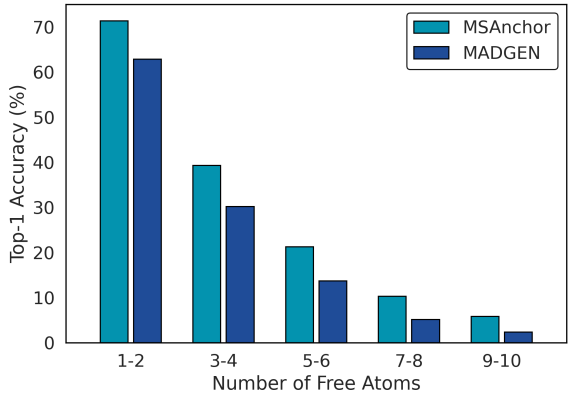

Figure 5: Performance comparison between MSAnchor and MADGEN under varying free atom counts.

NVIDIA A100 GPUs (40GB each). Further experimental settings and training cost are provided in Appendix.

Results

Table 1 presents the detailed experimental results on CANOPUS and MassSpecGym datasets using multiple evaluation metrics. On both datasets, MSAnchor achieves superior accuracy compared to all baseline models, demonstrating its superior capability in molecular structure prediction from mass spectral data. Specifically, on the CANOPUS dataset, most baseline models achieved partially correct predictions, while the MassSpecGym dataset to be proved significantly more challenging, with only the top three models achieving non-zero accuracy scores. This increased difficulty can be attributed to MassSpecGym's MCES-based dataset splitting strategy, which places higher demands on model generalization to novel molecular structures. MSAnchor's outstanding performance on MassSpecGym, with a top-1 accuracy of 2.68%, highlights its enhanced robustness in real-world scenarios. Furthermore, as shown in Figure 4, the consistent improvements in MCES and Tanimoto coefficient scores indicate that MSAnchor frequently generates structures with high structural similarity to ground truth molecules, even when exact matches are not achieved. Such structurally similar predictions remain highly valuable for downstream molecular discovery tasks, as demonstrated by the representative examples.

Sensitivity Analysis on Molecular Complexity

To investigate how molecular complexity influences the accuracy of structure completion, we evaluate the second-stage performance of MSAnchor in comparison with MADGEN, which also employs a two-step generation strategy. Specifically, we analyze the prediction accuracy on molecules with varying numbers of free atoms, defined as atoms not part of the scaffold. As illustrated in Figure 5, the top-1 accuracy of both models declines as the number of free atoms increases, reflecting the growing structural complexity and combinatorial challenge associated with higher numbers of free atoms. Notably, MSAnchor consistently outperforms MADGEN across all levels of complexity, with a particularly pronounced advantage in cases involving more free atoms. These results highlight MSAnchor's superior capacity to handle complex topological structures and accurately complete challenging molecular graphs.

Ablation study

Ablation on CIB To evaluate the impact of the CIB compression module, we conducted a comparison between MSAnchor and a variant without this component. As shown in Table 2, removing the CIB module leads to a slight drop in prediction performance, indicating that the module helps distill task-relevant information from the raw spectra while suppressing noise and redundant signals that could interfere with molecular reconstruction. Furthermore, in Appendix, we investigate the effects of different spectral encoding strategies on model performance. The results further support the design choice of employing both the current preprocessing pipeline and the CIB module, which together contribute to a more effective representation for structure generation.

Ablation on AEMS and Side Chain Priors To evaluate the impact of AEMS and side-chain priors, we com-

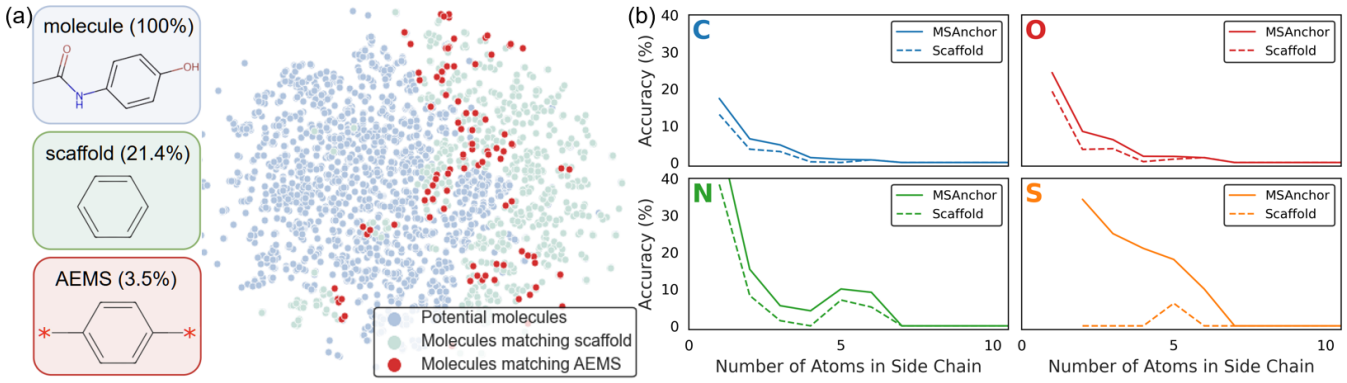

Figure 6: (a) Visualization of the molecular search space corresponding to a given molecular formula. Blue points denote all candidate molecules matching the formula, while green and red highlight subsets consistent with the molecular scaffold and AEMS predictions, respectively. (b) Comparison between MSAnchor and the variant using murcko scaffold on branch-level prediction accuracy across different atom types.

pare MSAnchor with a variant that uses the Murcko scaffold and omits attachment points and side-chain priors. As shown in Table 2, replacing AEMS with the Murcko scaffold significantly reduces prediction performance, highlighting AEMS's effectiveness in narrowing the search space and enabling fine-grained side-chain generation.To illustrate this limitation, we conducted a candidate search on PubChem using a given target molecule and analyzed the proportion of retrieved molecules that matched different scaffold types. As shown in Figure 6 (a), 21.4% of the retrieved candidates shared the same Murcko scaffold with the target, whereas only 3.5% matched the AEMS. This substantial reduction in match rate demonstrates the higher structural specificity and stricter constraint imposed by AEMS, which helps to significantly narrow down the search space for molecular generation and reduce ambiguity during reconstruction.

To further evaluate the effect of AEMS on fine-grained side-chain prediction, we analyzed model performance across branches with varying atom types and counts. Figure 6 (b) shows that MSAnchor consistently outperforms the comparison model, especially for branches containing nitrogen (N) and sulfur (S). The improved performance is likely due to the fact that these branches are relatively underrepresented in the dataset, which increases the difficulty of accurate reconstruction. The anchor-specific priors provide strong structural cues and help the model better capture and reconstruct such structures.

Ablation on Molecular Formula MSAnchor leverages both mass spectra and molecular formulas for structure prediction. While molecular formulas can be inferred from spectra using specialized tools (Barone et al. 2021; Valenti and Piskunov 1996), they are not always readily available in practice. To assess the impact of missing formulas, we adopt a two-stage strategy: first, candidate formulas are predicted using MIST-CF (Goldman et al. 2023b), then used as auxiliary input to MSAnchor. As shown in Table 2, using predicted formulas causes a moderate drop in accuracy compared to ground-truth input, likely due to errors in formula prediction. Nevertheless, MSAnchor is still able to generate structurally similar molecules, as the inclusion of chemically related formulas helps guide the model toward relevant candidates and enhances its robustness.

Model	Top-1		Top-10	
	Acc.↑	Sim.↑	Acc.↑	Sim.↑
w/o CIB	2.44%	0.29	4.31%	0.40
w/o AEMS	2.28%	0.27	4.17%	0.38
w/o Prior	2.35%	0.28	4.29%	0.39
w/o Formula	1.72%	0.26	4.04%	0.37
MSAnchor	**2.68%**	**0.32**	**4.67%**	**0.42**

Table 2: Ablation study on the MassSpecGym dataset.

Conclusion

In this work, we present MSAnchor, a novel framework for molecular structure prediction from mass spectra. At its core, MSAnchor introduces the Anchor-Extended Molecular Scaffold (AEMS) and employs a diffusion model to complete molecules based on AEMS. By explicitly specifying anchoring points for side chains, AEMS simplifies the molecular generation process and significantly reduces the combinatorial search space. Furthermore, MSAnchor ensures precise alignment between generated substructures and spectral features by leveraging spectrum-derived anchor-level priors to guide substructure generation, and employing a CIB module to extract informative and structure-relevant features from spectral data. Extensive experiments on two benchmark datasets demonstrate that MSAnchor achieves state-of-the-art performance, improving both prediction accuracy and interpretability. By explicitly modeling the attachment and composition of molecular branches, MSAnchor enables the generation of chemically valid molecules that align well with spectral characteristics, offering a reliable and interpretable approach for controllable molecular design.

Acknowledgements

The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Science. This paper is partially supported by the Project of Stable Support for Youth Team in Basic Research Field, CAS(YSBR-005), Natural Science Foundation of China National Major Research Instrument Development Project (No.12227901), the National Natural Science Foundation of China (No.62502491).

References

Aksenov, A. A.; da Silva, R.; Knight, R.; Lopes, N. P.; and Dorrestein, P. C. 2017. Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry*, 1(7): 0054.

Alberts, M.; Schilter, O.; Zipoli, F.; Hartrampf, N.; and Laino, T. 2024. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37: 125780–125808.

Barone, V.; Alessandrini, S.; Biczysko, M.; Cheeseman, J. R.; Clary, D. C.; McCoy, A. B.; DiRisio, R. J.; Neese, F.; Melosso, M.; and Puzzarini, C. 2021. Computational molecular spectroscopy. *Nature Reviews Methods Primers*, 1(1): 38.

Beniddir, M. A.; Kang, K. B.; Genta-Jouve, G.; Huber, F.; Rogers, S.; and Van Der Hooft, J. J. 2021. Advances in decomposing complex metabolite mixtures using substructure-and network-based computational metabolomics approaches. *Natural product reports*, 38(11): 1967–1993.

Blaženović, I.; Kind, T.; Ji, J.; and Fiehn, O. 2018. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites*, 8(2): 31.

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*.

Bristow, A. W. 2006. Accurate mass measurement for the determination of elemental formula—a tutorial. *Mass spectrometry reviews*, 25(1): 99–111.

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.

Coley, C. W. 2021. Defining and exploring chemical spaces. *Trends in Chemistry*, 3(2): 133–145.

de Hoffmann, E. 1996. Tandem mass spectrometry: a primer. *Journal of mass spectrometry*, 31(2): 129–137.

De Vijlder, T.; Valkenborg, D.; Lemière, F.; Romijn, E. P.; Laukens, K.; and Cuyckens, F. 2018. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass spectrometry reviews*, 37(5): 607–629.

Dean, J. C.; and Scholes, G. D. 2017. Coherence spectroscopy in the condensed phase: Insights into molecular structure, environment, and interactions. *Accounts of chemical research*, 50(11): 2746–2755.

Du, W.; Zhang, S.; Cai, Z.; Li, X.; Liu, Z.; Fang, J.; Wang, J.; Wang, X.; and Wang, Y. 2025. Molecular Merged Hypergraph Neural Network for Explainable Solvation Gibbs Free Energy Prediction. *Research*, 8: 0740.

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

Goldman, S.; Bradshaw, J.; Xin, J.; and Coley, C. 2023a. Prefix-tree decoding for predicting mass spectra from molecules. *Advances in neural information processing systems*, 36: 48548–48572.

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.

Hong, Y.; Ye, Y.; and Tang, H. 2025. Machine Learning in Small-Molecule Mass Spectrometry. *Annual Review of Analytical Chemistry*, 18.

Houhou, R.; and Bocklitz, T. 2021. Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. *Analytical Science Advances*, 2(3-4): 128–141.

Hu, W.; Liu, Y.; Chen, X.; Chai, W.; Chen, H.; Wang, H.; and Wang, G. 2023. Deep learning methods for small molecule drug discovery: A survey. *IEEE Transactions on Artificial Intelligence*, 5(2): 459–479.

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.

Kretschmer, F.; Seipp, J.; Ludwig, M.; Klau, G. W.; and Böcker, S. 2023. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv*.

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

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.

Liyaqat, T.; Ahmad, T.; and Saxena, C. 2024. Advancements in molecular property prediction: a survey of single and multimodal approaches. *arXiv preprint arXiv:2408.09461*.

Luo, Y.; Fang, J.; Li, S.; Liu, Z.; Wu, J.; Zhang, A.; Du, W.; and Wang, X. 2024. Text-guided Diffusion Model for 3D Molecule Generation. arXiv:2410.03803.

Ma, K.; Vitek, O.; and Nesvizhskii, A. I. 2012. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC bioinformatics*, 13(Suppl 16): S1.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

McLafferty, F. W. 1981. Tandem mass spectrometry. *Science*, 214(4518): 280–287.

McLuckey, S. A.; and Stephenson, J. L. 1998. Ion/ion chemistry of high-mass multiply charged ions. *Mass spectrometry reviews*, 17(6): 369–407.

Morgan, H. L. 1965. 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.

Nguyen, D. H.; Nguyen, C. H.; and Mamitsuka, H. 2019. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Briefings in bioinformatics*, 20(6): 2028–2043.

Niessen, W. M.; et al. 2017. *Interpretation of MS-MS mass spectra of drugs and pesticides*. John Wiley & Sons.

Oprea, T. I.; and Gottfries, J. 2001. Chemography: the art of navigating in chemical space. *Journal of combinatorial chemistry*, 3(2): 157–166.

Ross, J.; Belgodere, B.; Hoffman, S. C.; Chenthamarakshan, V.; Navratil, J.; Mroueh, Y.; and Das, P. 2024. Gp-molformer: A foundation model for molecular generation. *arXiv preprint arXiv:2405.04912*.

Scheubert, K.; Hufsky, F.; and Böcker, S. 2013. Computational mass spectrometry for small molecules. *Journal of cheminformatics*, 5(1): 12.

Sleno, L.; and Volmer, D. A. 2004. Ion activation methods for tandem mass spectrometry. *Journal of mass spectrometry*, 39(10): 1091–1112.

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

Valenti, J. A.; and Piskunov, N. 1996. Spectroscopy made easy: A new tool for fitting observations with synthetic spectra. *Astronomy and Astrophysics Supplement Series*, 118(3): 595–603.

Vaniya, A.; and Fiehn, O. 2015. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *TrAC Trends in Analytical Chemistry*, 69: 52–61.

Vestal, M. L. 2001. Methods of ion generation. *Chemical reviews*, 101(2): 361–376.

Walters, W. P.; and Barzilay, R. 2020. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2): 263–270.

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, Z.; and Barati Farimani, A. 2023. Graph neural networks for molecules. In *Machine learning in molecular sciences*, 21–66. Springer.

Winger, B. E.; Light-Wahl, K. J.; Ogorzalek Loo, R. R.; Udseth, H. R.; and Smith, R. D. 1993. Observation and implications of high mass-to-charge ratio ions from electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 4(7): 536–545.

Wishart, D. S. 2016. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature reviews Drug discovery*, 15(7): 473–484.

Xia, J.; Zhang, L.; Zhu, X.; Liu, Y.; Gao, Z.; Hu, B.; Tan, C.; Zheng, J.; Li, S.; and Li, S. Z. 2023. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. *Advances in Neural Information Processing Systems*, 36: 64774–64792.

Xue, X.; Sun, H.; Yang, M.; Liu, X.; Hu, H.-Y.; Deng, Y.; and Wang, X. 2023. Advances in the application of artificial intelligence-based spectral data interpretation: a perspective. *Analytical chemistry*, 95(37): 13733–13745.

Zhang, S.; Fang, J.; Li, X.; hongxin xiang; XIA, A.; Wei, Y.; Du, W.; and Wang, Y. 2025. Iterative Substructure Extraction for Molecular Relational Learning with Interactive Graph Information Bottleneck. In *The Thirteenth International Conference on Learning Representations*.

Zhang, X.; Wang, S.; Zhu, F.; Xu, Z.; Wang, Y.; and Huang, J. 2018. Seq3seq fingerprint: towards end-to-end semi-supervised deep drug discovery. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 404–413.