De Novo Molecular Generation from Mass Spectra via Many-Body Enhanced Diffusion

Xichen Sun[1,3*], **Wentao Wei**[1,2*], **Jiahua Rao**[1†], **Jiancong Xie**[1], **Yuedong Yang**[1,4†]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China
[2]Pengcheng Laboratory, Shenzhen, Guangdong, China
[3]Shenzhen Loop Area Institute, Shenzhen, Guangdong, China
[4]Guangdong Provincial Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, Guangdong, China
{sunxch7, weiwt8, xiejc3}@mail2.sysu.edu.cn, {raojh7, yangyd25}@mail.sysu.edu.cn

Abstract

Molecular structure generation from mass spectrometry is fundamental for understanding cellular metabolism and discovering novel compounds. Although tandem mass spectrometry (MS/MS) enables the high-throughput acquisition of fragment fingerprints, these spectra often reflect higher-order interactions involving the concerted cleavage of multiple atoms and bonds-crucial for resolving complex isomers and non-local fragmentation mechanisms. However, most existing methods adopt atom-centric and pairwise interaction modeling, overlooking higher-order edge interactions and lacking the capacity to systematically capture essential many-body characteristics for structure generation. To overcome these limitations, we present **MBGen**, a **M**any-**B**ody enhanced diffusion framework for de novo molecular structure **Gen**eration from mass spectra. By integrating a many-body attention mechanism and higher-order edge modeling, MBGen comprehensively leverages the rich structural information encoded in MS/MS spectra, enabling accurate de novo generation and isomer differentiation for novel molecules. Experimental results on the NPLIB1 and MassSpecGym benchmarks demonstrate that MBGen achieves superior performance, with improvements of up to 230% over state-of-the-art methods, highlighting the scientific value and practical utility of many-body modeling for mass spectrometry-based molecular generation. Further analysis and ablation studies show that our approach effectively captures higher-order interactions and exhibits enhanced sensitivity to complex isomeric and non-local fragmentation information.

Code — https://github.com/biomed-AI/MBGen

Introduction

The comprehensive understanding of cellular metabolism is essential for advancing basic biological research and applied biomedical sciences (Wishart 2019; DeBerardinis and Thompson 2012; Rao et al. 2025a; Xie et al. 2024). Metabolomics, which focuses on the systematic profiling of small molecules in biological samples, plays a pivotal role in revealing metabolic pathways and disease mechanisms (Vander Heiden, Cantley, and Thompson 2009; Qiu et al. 2023; Newgard 2017; Rao et al. 2024). At the heart of metabolomics lies tandem mass spectrometry (MS/MS), a powerful analytical technique that enables high-throughput detection and structural elucidation of diverse metabolites through detailed fragmentation spectra (Kind et al. 2018). These spectra provide molecular fingerprints that enable the annotation of known compounds and the discovery of novel ones. The widespread adoption of MS/MS has markedly improved the resolution and depth of metabolomic analyses, propelling innovations in the field.

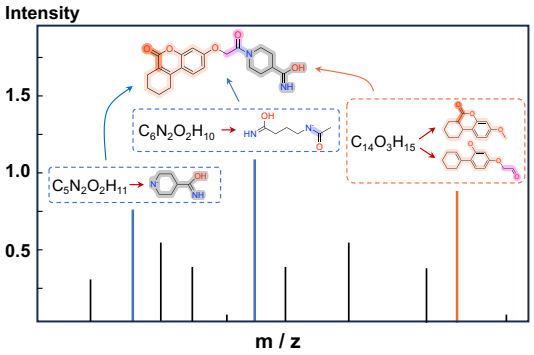

Figure 1: The mass spectrum comprises fragments from complex bond cleavages, where isomers within the same peak each contribute distinct yet essential insights for accurate molecular generation.

Despite its transformative impact, accurate and automated molecular structure generation from MS/MS data remains a fundamental challenge. The fragmentation spectra produced by tandem mass spectrometry are inherently complex, often reflecting not only the cleavage of individual chemical bonds but also higher-order interactions involving the concerted breakage of multiple atoms and bonds. These many-body effects encode rich structural information crucial for distinguishing complex isomers and interpreting non-local fragmentation mechanisms (Tanaka and Arita 2018).

*These authors contributed equally.
†Corresponding Author.

Existing computational approaches for molecular structure generation from MS/MS data primarily adopt an atom-centric perspective. For example, language models that map tokenized m/z values and intensities to SMILES strings (Litsa et al. 2023; Stravs et al. 2022; Winter et al. 2019), as well as graph-based methods like MADGEN (Wang et al. 2025) and DiffMS (Bohde et al. 2025), represent and generate molecular structures by focusing on atoms as fundamental units. While these atom-centric strategies have improved prediction accuracy and enabled basic molecular assembly, they inherently overlook the rich chemical information encoded in bonds and the interactions between them. However, bond formation and cleavage events are central to the fragmentation processes observed in MS/MS spectra. Accurately modeling these processes requires a representation that directly captures the connectivity and dynamics of chemical bonds, rather than merely the arrangement of atoms.

Moreover, the fragmentation patterns observed in the MS/MS data often result from the concerted breaking of multiple bonds and complex, interdependent interactions that cannot be captured by simple pairwise modeling. Even advanced GNN-based models, such as those employing the Graph Transformer in DiffMS (Bohde et al. 2025), are limited in their ability to encode such higher-order relationships. Explicitly incorporating many-body interactions is therefore essential, as it enables the model to represent and predict multi-bond cleavage dynamics and non-local fragmentation mechanisms that are critical for faithful interpretation of MS/MS spectra. By capturing these complex interactions, the model can more effectively resolve structural isomers and generate chemically plausible de novo structures, ultimately leading to more accurate molecular generation from spectral data.

In this work, we present **MBGen**, a **M**any-**B**ody enhanced diffusion framework for de novo molecular **Gen**eration from mass spectra. Specifically, MBGen differs from traditional atom-centric frameworks by adopting an edge-centric molecular generation strategy, modeling molecules at the level of chemical connectivity. This approach provides a more direct and chemically meaningful connection between spectral features and molecular structure, since bond formation and cleavage are fundamental to MS/MS fragmentation. Furthermore, MBGen incorporates a many-body attention mechanism to explicitly capture higher-order interactions and concerted bond-breaking events. This allows the model to learn complex fragmentation pathways beyond simple pairwise or conventional GNN-based modeling, enabling more accurate interpretation of MS/MS spectra and effective differentiation of structural isomers. By integrating these chemical insights into a diffusion-based generative process, MBGen flexibly and reliably generates chemically plausible molecular structures from spectral data, advancing de novo molecular generation and isomer identification.

Experimental results on the NPLIB1 (Dührkop et al. 2021) and MassSpecGym (Bushuiev et al. 2024) benchmarks demonstrate that MBGen achieves superior performance, with improvements of up to 230% over state-of-the-art methods. MBGen consistently outperforms baseline models in distinguishing structural isomers, particularly in cases where isomeric species produce highly similar fragmentation patterns. Further analysis and ablation studies confirm that our approach effectively captures higher-order interactions and exhibits enhanced sensitivity to complex isomeric and non-local fragmentation information. These strengths position MBGen as a valuable tool for applications such as metabolite identification, drug discovery, and the structural elucidation of novel compounds in complex biological samples. Our main contributions are as follows:

- We adopt an edge-centric molecular modeling strategy that directly represents chemical bonds and their connectivity, providing a more accurate foundation for interpreting MS/MS fragmentation.
- We incorporate a many-body attention mechanism throughout the molecular generation process, explicitly capturing higher-order interactions and concerted bond-breaking events, which enables the model to better resolve complex fragmentation and structural isomers.
- Extensive experiments demonstrate that MBGen significantly outperforms existing methods in both molecular structure generation accuracy and isomer differentiation, validating the value and practical utility of our approach.

Related Work

Molecular Generation based on Mass Spectra

Identifying molecular structures from mass spectrometry (MS) data remains challenging. Traditional approaches, such as those by Heinonen et al. (2012) and tools like CSI:FingerID (Dührkop et al. 2015), predict molecular properties or fingerprints from tandem MS spectra and match them against molecular databases. However, these database-dependent workflows are computationally intensive and fundamentally limited by the coverage of reference databases, making the identification of novel compounds impossible.

To overcome these limitations, de novo molecular generation methods have been developed. These approaches predict molecular structures directly from MS data. For example, MSNovelist (Stravs et al. 2022) combines predicted molecular fingerprints and formulas with an autoregressive model for molecule reconstruction. Spec2Mol (Litsa et al. 2023) employs an encoder-decoder framework, mapping spectra into a learned molecular embedding space for structure generation. MADGEN (Wang et al. 2025) uses a two-stage process: scaffold retrieval from spectra, followed by conditional structure generation. DiffMS (Bohde et al. 2025) further advances the field with an end-to-end framework, pretraining spectra and structure modules separately before joint finetuning. However, these methods typically adopt atom-centric and pairwise interaction modeling, overlooking higher-order edge interactions and lacking the capacity to capture essential many-body characteristics.

Many-body Interaction Modeling

Recent advances in molecular modeling have highlighted the importance of many-body interactions for capturing complex dependencies between atoms in a molecule. While traditional geometric GNNs (Ying et al. 2021; Hussain, Zaki,

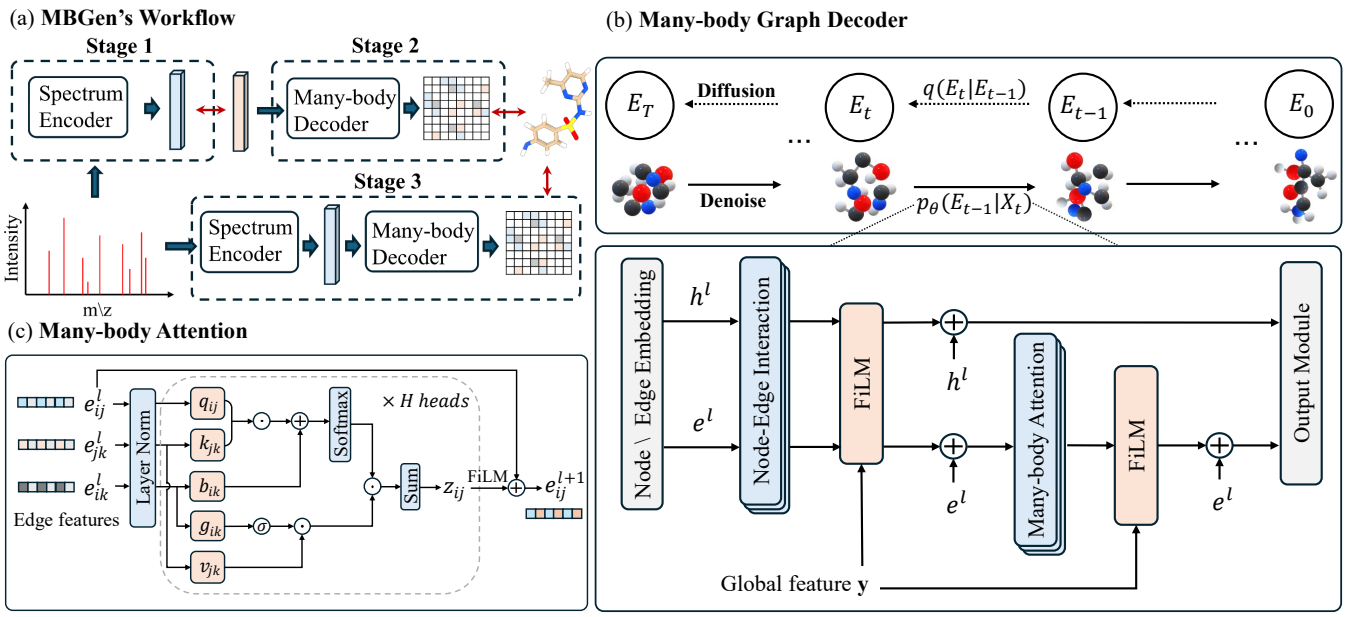

Figure 2: An overview of MBGen framework.

and Subramanian 2022; Rao et al. 2025b) mainly focus on pairwise interactions, limiting their expressiveness, recent architectures such as ViSNet (Wang et al. 2024), GEM-2 (Liu et al. 2022), TGT (Hussain, Zaki, and Subramanian 2024), and MABNet (Rao et al. 2025c) have incorporated higher-order mechanisms-modeling three-body and even four-body interactions-to enhance molecular representation learning. However, these approaches have largely been restricted to property prediction tasks, with limited exploration in molecular generation.

In this work, we bridge this gap by introducing many-body interaction modeling directly into the generative process, enabling the decoder to capture richer geometric relationships and thereby improving the fidelity and accuracy of generated molecular structures.

Methodology

Problem Formulation

We formulate molecular structure generation from mass spectrometry (MS) data as a conditional graph generation problem. A molecule is represented as a graph $G = (H, E)$, where $H \in \mathbb{R}^{n \times d}$, $E \in \{0,1\}^{n \times n \times k}$, with n heavy atoms, d feature dimensions, and k bond types.

Given a chemical formula which is obtainable from high-resolution MS data using tools like SIRIUS (Böcker and Dührkop 2016), the node set H is fixed, and the task reduces to predicting E so that the resulting graph matches the observed spectrum. The input spectrum S is comprised of m/z peaks and intensities I, which are encoded into a vector y that provides global structural constraints. Under this setting, our model learns to predict the adjacency matrix E conditioned on the fixed node features H and the spectral embedding y.

Model Architecture

Figure 2 illustrates our overall framework. The input is a tandem mass spectrum, represented as a set of $(m/z, \text{intensity})$ peaks. The model first employs a spectrum encoder to extract a structural fingerprint from the spectrum. This fingerprint conditions a many-body graph diffusion decoder, which iteratively denoises a molecular graph to generate the final structure. The decoding process incorporates an edge-centric message passing mechanism and a many-body attention module to capture rich relationships and higher-order structural contexts. Our model adopts a three-stage training procedure, consisting of spectrum encoder pretraining, graph decoder pretraining, and end-to-end finetuning.

Spectrum Encoder

Following previous work (DiffMS), we use the pretrained MIST (Goldman et al. 2023) Formula Transformer as our spectrum encoder to extract molecular fingerprints from tandem mass spectra. Given an input spectrum $S = \{(m/z_i, \text{intensity}_i)\}_{i=1}^{N}$, the encoder maps S to a fixed-dimensional fingerprint vector y:

$$y = \text{Encoder}(S), \tag{1}$$

where $y \in \mathbb{R}^d$ serves as the structural representation conditioning the subsequent molecular graph generation.

For the encoder, we first apply SIRIUS to annotate each peak with its most probable molecular formula, which is then concatenated with the corresponding peak intensity:

$$x_i = [F_i; \text{intensity}_i], \tag{2}$$

where F_i embeds the molecular formula with a learned formula embedding function.

The set of peaks $X = \{x_i\}_{i=1}^{N}$ is then encoded by a Set Transformer comprising pairwise attention layers. Each

layer updates peak embeddings by modeling their interactions as:

$$\mathrm{Attn}(x_i, x_j) = \frac{(Q_i + b_1)K_j + (Q_i + b_2)|F_i - F_j|}{\sqrt{d}}, \quad (3)$$

where Q_i, K_j are the query and key vectors from learned projections of x_i, x_j, and b_1, b_2 are trainable bias terms. The value $|F_i - F_j|$ denotes the element-wise absolute difference between the embedded formulas of peaks i and j.

Finally, the molecular fingerprint y is obtained via mean pooling over the encoded representations:

$$y = \frac{1}{N}\sum_{i=1}^{N} x_i. \quad (4)$$

Many-body Enhanced Graph Decoder

Edge-centric Molecular Modeling. For decoder, we adopt an edge-centric strategy, where information propagation and representation are centered on chemical bonds (edges) rather than solely on atoms (nodes). Instead of focusing exclusively on node-level embeddings, we explicitly construct edge features based on the associated node features and their relational context.

We first initialize node features for each atom i as:

$$h_i^{(0)} = \mathrm{NodeEmb}(a_i), \quad (5)$$

where a_i encodes the atomic type and intrinsic properties. Then, the pairwise (edge) embedding between nodes i and j is constructed as:

$$e_{ij}^{(0)} = \mathrm{EdgeEmb}\big(h_i^{(0)}, h_j^{(0)}, r_{ij}\big), \quad (6)$$

where r_{ij} represents the relationship between i and j.

In the edge-centric message passing, edge embeddings are updated in two sequential stages. First, node-edge interaction layers aggregate information from associated node features and the global feature y to refine edge embeddings:

$$e_{ij}^{(l+1)} = f_{\mathrm{Edge}}\Big(e_{ij}^{(l)}, h_i^{(l)}, h_j^{(l)}, y\Big). \quad (7)$$

Formally, at each layer l, the attention-weighted aggregation of node and edge features is computed as follows:

$$\alpha_{ij}^{(l)} = \frac{(W_Q \cdot h_i^{(l)})(W_K \cdot h_j^{(l)})^\top}{\sqrt{d}} + W_E \cdot e_{ij}^{(l)}, \quad (8)$$

$$o_{ij}^{(l)} = W_\alpha \cdot \alpha_{ij}^{(l)}, \quad (9)$$

where W_Q, W_K, W_E and W_α are learnable projection matrices. Here, $o_{ij}^{(l)}$ represents the intermediate attention result integrating node and edge information.

Then, we apply a FiLM (Perez et al. 2018) mechanism to incorporate the global feature y into edge representations. The modulation is applied as:

$$\mathrm{FiLM}(o_{ij}^{(l)}, y) = yW_2 + o_{ij}^{(l)} \cdot (yW_1) + o_{ij}^{(l)}, \quad (10)$$

where W_1, W_2 are linear transformations.

Finally, the edge embedding is updated through a feed-forward network FFN_e:

$$e_{ij}^{(l+1)} = \mathrm{FFN}_e\left(e_{ij}^{(l)} + \mathrm{FiLM}(o_{ij}^{(l)}, y)\right). \quad (11)$$

After updating edge features via node-edge interaction, to further capture higher-order chemical interactions, we incorporate a many-body attention mechanism, denoted as f_{ManyBody}. Specifically, for each pair (i, j), the many-body attention updates the pairwise embedding e_{ij} by aggregating information not only from nodes i and j, but also considering all neighbor pairs (j, k) and their relationship with (i, k), thus capturing interaction patterns among triplets (i, j, k):

$$e_{ij}^{(l+1)} = f_{\mathrm{ManyBody}}\Big(e_{ij}^{(l)}, \{e_{ik}^{(l)}, e_{jk}^{(l)} \mid k \in \mathcal{N}(j)\}, y\Big), \quad (12)$$

where $\mathcal{N}(j)$ denotes the neighborhood of node j. This enables the model to encode richer structural context by explicitly modeling interactions among triplets (i, j, k), which is crucial for capturing complex fragmentation and higher-order chemical relationships during molecular generation.

Many-body Attention Module. As shown in Figure 2(b), after the node-edge interaction, the pairwise embedding e_{ij} is updated by a many-body attention module, which is further illustrated in detail in Figure 2(c). Specifically, at layer l, the many-body attention computes an intermediate output $\mathbf{z}_{ij}^{(l)}$ for each pair (i, j) by performing a weighted sum over the value vectors of neighboring pairs:

$$\mathbf{z}_{ij}^{(l)} = \sum_{k=1}^{N} \alpha_{ijk}\, \mathbf{v}_{jk} \quad (13)$$

where the attention weight α_{ijk} reflects the contribution of the neighbor pair (j, k) when updating the target pair (i, j), and v_{jk} denotes the value vector of pair embedding $e_{jk}^{(l)}$. Specifically, α_{ijk} is computed as follows:

$$s_{ijk} = \frac{1}{\sqrt{d}}\, \mathbf{q}_{ij} \cdot \mathbf{k}_{jk} + b_{ik}, \quad (14)$$

$$\alpha_{ijk} = \mathrm{softmax}_k(s_{ijk}) \cdot \sigma(g_{ik}). \quad (15)$$

Here, the query, key, and value vectors are computed as linear projections of the corresponding pairwise embeddings:

$$\mathbf{q}_{ij} = W_Q \cdot e_{ij}^{(l)}, \mathbf{k}_{jk} = W_K \cdot e_{jk}^{(l)}, \mathbf{v}_{jk} = W_V \cdot e_{jk}^{(l)}, \quad (16)$$

and the bias term b_{ik} and gating vector g_{ik} are similarly computed from the third pairwise embedding $e_{ik}^{(l)}$:

$$g_{ik} = W_G \cdot e_{ik}^{(l)}, b_{ik} = W_B \cdot e_{ik}^{(l)}, \quad (17)$$

where all projection matrices W_Q, W_K, W_V, W_G and W_B are learned parameters.

The final attention weight is obtained by applying a softmax function and modulating it with a sigmoid gate $\sigma(g_{ik})$, which adaptively filters irrelevant interactions.

Similar to the previous step, we incorporate the global feature y using the same FiLM method:

$$\mathrm{FiLM}(z_{ij}^{(l)}, y) = yW_2 + z_{ij}^{(l)} \cdot (yW_1) + z_{ij}^{(l)}. \quad (18)$$

The final updated embedding is then computed as:

$$e_{ij}^{(l+1)} = e_{ij}^{(l)} + \mathrm{FiLM}(z_{ij}^{(l)}, y). \quad (19)$$

This many-body update enables information flow along triplets without necessarily involving the junction node j, effectively alleviating the bottleneck and elevating the model's expressivity. A detailed analysis of computational efficiency is provided in Appendix E.

Discrete Diffusion

As shown in Figure 2(b), the discrete diffusion generation involves two processes: i) A diffusion process gradually corrupts the edge features of the molecular graph by introducing discrete noise; ii) A denoising process learns to reconstruct the molecular graph conditioned on spectral embeddings.

Diffusion Process. Given a structure-spectrum pair $X = (G, y)$, where $G = (H, E)$ is a molecular graph and $E \in \mathbb{R}^{n\times n\times k}$ denotes the edge feature, we represent each edge feature e as a k-dimensional one-hot vector, with class 0 being non-edge and classes 1 to k-1 corresponding to different bond types. We model its diffusion process via a discrete forward process over time steps $t = 0, 1, \ldots, T$, progressively adding noise to edge features using a categorical transition matrix Q_t:

$$q(E_t|E_{t-1}) = E_{t-1}Q_t \quad \text{and} \quad q(E_t|E) = E\bar{\mathbf{Q}}_\mathbf{t}, \quad (20)$$

where $Q_t = [q(e_t = j \mid e_{t-1} = i)]_{i,j=0}^{k-1}$ and $\bar{\mathbf{Q}}_t = Q_tQ_{t-1}\cdots Q_1$.

For undirected graphs, noise is applied to the upper-triangular part of E, which is then symmetrized.

Denoising Process. The reverse process begins with the fully corrupted edge matrix E_T and iteratively generates E_{t-1} from E_t until E_0, aiming to reconstruct the original edge types. A neural network f_θ is trained to directly estimate E_0 from E_t, conditioned on the noisy graph and the global feature. Consequently, the denoising transition from E_t to E_{t-1} can be expressed as:

$$q(E_{t-1} \mid E_t, E_0) \approx p_\theta(E_{t-1} \mid X_t), \quad (21)$$

where θ serves as a learnable parameter.

Modeling the Denoising Network f_θ. To preserve the information of the molecular fragments interactions, we utilize the many-body attention mechanism to model f_θ. The denoising neural network f_θ is trained to reverse the corruption by predicting $\hat{p} = f_\theta(G_t, t, y)$. f_θ takes a noisy graph G_t as input and aims to predict the clean graph G. To train f_θ, we optimize the cross-entropy loss L between the predicted probabilities $\hat{p}$ and the true edge matrix E:

$$L = \mathrm{CE}(\hat{p}, E). \quad (22)$$

After obtaining the trained network f_θ, new graphs are generated by estimating the reverse diffusion iterations $p_\theta(E_{t-1}|X_t)$, using marginalization over the network's predicted distribution $\hat{p}$:

$$p_\theta(e_{ij}^{t-1}|X_t) = \sum_{e\in E} p_\theta(e_{ij}^{t-1}|e_{ij} = e, X_t)\hat{p}(e), \quad (23)$$

where $p_\theta(e_{ij}^{t-1}|e_{ij} = e, X_t) = \max\Big(q(e_{ij}^{t-1}|e_{ij} = e, e_{ij}^t), 0\Big)$. We can sample a discrete E_{t-1} from the distribution and then use it as the input for the denoising network in the next time step, iteratively performing the diffusion process until E_0.

Training Paradigm

Spectrum Encoder Pretraining. Following the strategy of DiffMS (Bohde et al. 2025), we first pretrain the spectrum encoder on NPLIB1 and MassSpecGym datasets. Given an input spectrum S, the encoder is trained to predict the corresponding molecular fingerprint. This stage encourages the encoder to extract structural signals from spectral data and provides a strong starting point for subsequent finetuning.

Many-body Decoder Pretraining. To enhance the decoder's ability to generate molecular structures under strong structural constraints, we pretrain the decoder independently on a large collection of molecular fingerprint–structure pairs. Specifically, instead of using spectrum embeddings as input, we directly use the molecular fingerprint as the structural condition y. The decoder is then trained to reconstruct the molecular graph, learning to generate true structures guided by the fingerprint feature.

End-to-End Finetuning. After separate pretraining of the encoder and decoder, we jointly finetune the entire model using the same datasets as encoder pretraining. The encoder processes input spectra to produce the fingerprint embeddings y, which are then used to condition the many-body decoder. In this stage, the model is trained to reconstruct the full molecular graph from scratch, aligning the predicted adjacency matrix with the ground-truth structure.

Experiments

Experiment Setup

In this section, we briefly describe the datasets, baseline methods, and evaluation metrics used in our experiments. Additional information can be found in Appendix A-C.

- **Datasets.** We pretrain the decoder on a large scale of 2.8 million fingerprint–molecule pairs collected from DSSTox (CCTE, EPA 2019), HMDB (Wishart et al. 2022), COCONUT (Sorokina et al. 2021), and MOSES (Polykovskiy et al. 2020), covering diverse chemical structures. Evaluation is conducted on two public benchmarks: NPLIB1 (Dührkop et al. 2021), and MassSpecGym (Bushuiev et al. 2024).
- **Baselines.** We compare our method with several state-of-the-art approaches, including Spec2Mol (Litsa et al. 2023), MIST (Goldman et al. 2023) combined with Neuraldecipher (Winter et al. 2019) or MSNovelist (Stravs et al. 2022), DiffMS (Bohde et al. 2025), and MADGEN (Wang et al. 2025). We also reproduce the results of SMILES Transformer, SELFIES Transformer, and Random Chemical Generation from MassSpecGym.
- **Evaluation Metrics.** Following MassSpecGym, we report top-k accuracy, Tanimoto similarity, and Maximum Common Edge Substructure (MCES) scores for $k = 1$ and $k = 10$. All models generate 100 molecular candidates for each spectrum.

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
MBGen	**12.20%**	**7.72**	**0.41**	**22.29%**	**6.71**	**0.50**
MassSpecGym						
SMILES Transformer	0.00%	79.39	0.03	0.00%	52.13	0.10
MIST + MSNovelist	0.00%	45.55	0.06	0.00%	30.13	0.15
SELFIES Transformer	0.00%	38.88	0.08	0.00%	26.87	0.13
Spec2Mol	0.00%	37.76	0.12	0.00%	29.40	0.16
MIST + Neuraldecipher	0.00%	33.19	0.14	0.00%	31.89	0.16
Random Chemical Generation	0.00%	21.11	0.08	0.00%	18.26	0.11
MADGEN	1.31%	27.47	0.20	1.54%	16.84	0.26
DiffMS	2.30%	18.45	0.28	4.25%	14.73	0.39
MBGen	**7.58%**	**13.25**	**0.38**	**12.54%**	**10.16**	**0.47**

Table 1: Evaluation of de novo molecular structure elucidation models on the NPLIB1 (Dührkop et al. 2021) and MassSpecGym (Bushuiev et al. 2024) datasets. The table presents top-1 and top-10 accuracy, Maximum Common Edge Substructure(MCES) scores, and Tanimoto similarity. **Bold** indicates the best performance.

Main results

Table 1 summarizes the main results of our method and representative baselines on the NPLIB1 and MassSpecGym datasets. Our method achieves state-of-the-art performance across all metrics. On NPLIB1, our approach reaches a Top-1 accuracy of 12.20%, surpassing the previous best result (DiffMS, 8.34%) by a substantial margin. Similarly, on MassSpecGym, our method attains a Top-1 accuracy of 7.58%, compared to 2.30% for DiffMS. Significant improvements are also observed in Top-10 accuracy, MCES, and Tanimoto similarity, indicating that our model not only predicts more accurate structures but also generates candidates with higher substructural and overall molecular similarity.

Despite the use of chemical formula constraints, existing models such as DiffMS still exhibit relatively limited accuracy. Our results demonstrate that explicitly modeling bond–bond (edge–edge) interactions, inspired by chemical fragmentation principles, can substantially improve both the accuracy and chemical plausibility of structure generation from MS/MS spectra. This highlights the importance of incorporating chemical knowledge into deep generative frameworks for molecular elucidation.

Pretrain		Metrics		
Enc.	Dec.	Accuracy ↑	MCES ↓	Tanimoto ↑
Top-1				
✗	✗	0.00%	17.96	0.17
✓	✗	4.17%	15.73	0.26
✗	✓	8.33%	15.20	0.30
✓	✓	**12.20%**	**7.72**	**0.41**
Top-10				
✗	✗	2.08%	14.26	0.25
✓	✗	8.33%	13.25	0.36
✗	✓	16.67%	11.65	0.44
✓	✓	**22.29%**	**6.71**	**0.50**

Table 2: Performance on the NPLIB1 dataset with and without pretraining of encoder and decoder.

Ablation Study

Many-body attention module. To evaluate the efficacy of the many-body attention module, we conduct ablation experiments by removing it from the model (denoted as w/o MB). We use MAGMA to annotate fragment molecules for each peak in the mass spectra and compare model performance across varying numbers of isomers. As illustrated in Figure 4(a-b), both MBGen and the ablated model perform similarly when the average isomer count per peak is less than 1, but as the isomer complexity increases, MBGen maintains stable performance while the ablated model degrades, highlighting the module's effectiveness.

We also assess the reconstruction of complex molecules (Fig 4(c-d)). For molecules with fewer than 23 atoms, both models perform comparably, but as the atom count increases, MBGen shows substantial gains. Notably, for molecules with over 40 atoms, MBGen achieves a Tanimoto similarity of about 0.525 versus 0.425 for the ablated model. Incorporating higher-order interactions thus enables robust modeling of complex molecules and offers deeper insights into molecular generation.

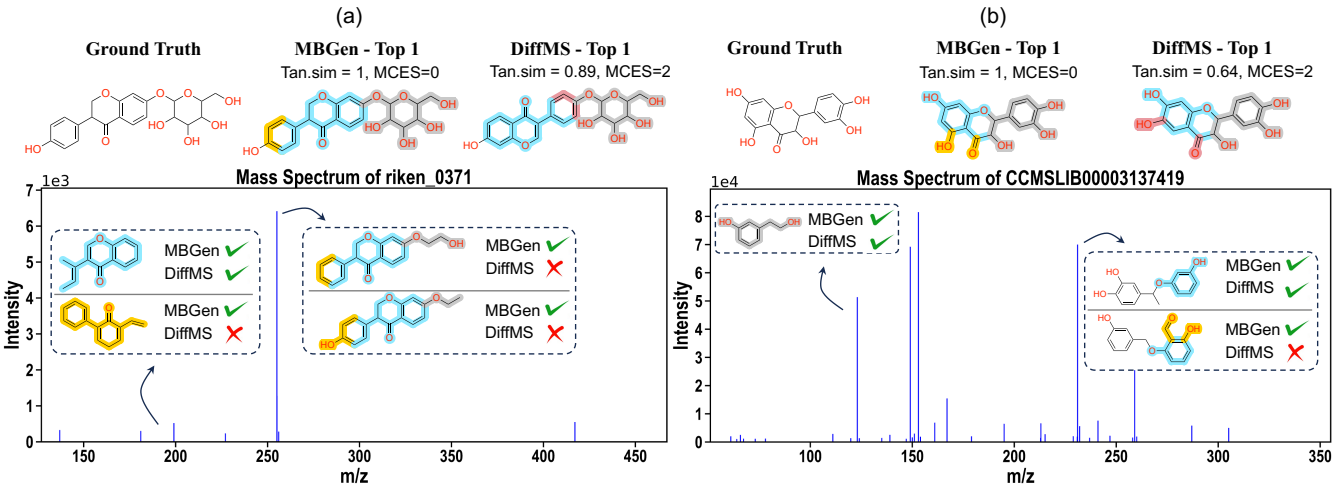

Figure 3: Case studies illustrating the superior performance of MBGen over DiffMS in de novo molecular structure generation, particularly in spectra featuring intra-peak isomers.

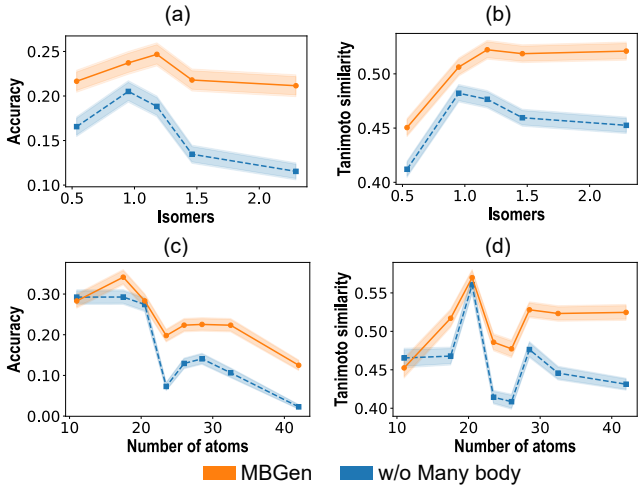

Figure 4: Ablation study on the NPLIB1 dataset, evaluating model performance across varying numbers of isomers and atoms.

Pretrain-finetune strategy. To assess the impact of our pretraining-finetuning strategy, we ablate the pretraining of the encoder and decoder components on the NPLIB1 dataset. As shown in Table 2, models without any pretraining yield poor performance, with Top-1 accuracy at 0.00% and Tanimoto similarity at 0.17. Pretraining only the encoder or decoder provides moderate gains (e.g., Top-1 accuracy of 4.17% and 8.33%, respectively), while pretraining both achieves the best results: Top-1 accuracy of 12.20%, MCES of 7.72 (lower is better), and Tanimoto of 0.41. Similar trends hold for Top-10 metrics, with the full strategy reaching 22.29% accuracy, 6.71 MCES, and 0.50 Tanimoto. These results demonstrate that joint pretraining of encoder and decoder significantly enhances de novo generation by leveraging spectral and molecular priors.

Case Study

To demonstrate the interpretability and efficacy of MBGen, we present case studies highlighting its ability to capture isomer information within mass spectral peaks, leading to accurate de novo molecular structure generation where baselines like DiffMS falter. As shown in Figure 3(a), for the mass spectrum of riken-0371, both MBGen and DiffMS capture the chromone structural motif among isomers at the peak m/z 199.08. However, MBGen additionally identifies critical isomer fragments related to the benzene ring and chromone positional arrangements, which DiffMS misses. For the more complex isomer fragments at peak m/z 255.07, MBGen successfully captures them, whereas DiffMS fails, resulting in the loss of key structural insights.

Similarly, Figure 3(b) illustrates the spectrum of CCMSLIB00003137419, where MBGen accurately reconstructs the molecule by leveraging the many-body interaction to integrate complementary isomer details across peaks, whereas DiffMS produces an incorrect structure. This underscores how MBGen's many-body algorithm enhances interpretability by explicitly accounting for intra-peak isomer complexity, yielding superior predictions on challenging cases.

Conclusion

In this work, we propose MBGen, a many-body enhanced diffusion framework with edge-centric modeling for de novo molecular generation from mass spectra. We develop a pretraining-finetuning workflow incorporating an edge-enhanced transformer and a many-body attention module that leverages higher-order bond interactions and intra-peak isomer information, ensuring the model captures chemically nuanced representations from spectral data. We show that MBGen achieves state-of-the-art results across de novo generation benchmarks, and provide ablation studies and case analyses to demonstrate the effectiveness of our contributions and the potential to further enhance performance by scaling pretraining or integrating additional spectral priors.

Acknowledgments

This study has been supported by Shenzhen Medical Research Fund [C2403001], the Guangdong S&T Program [2024B1111140001], the China Postdoctoral Science Foundation [2025M771540, GZB20250391], and the Lingang Laboratory [LGL-8888].

References

Böcker, S.; and Dührkop, K. 2016. Fragmentation trees reloaded. *Journal of cheminformatics*, 8(1): 5.

Bohde, M.; Manjrekar, M.; Wang, R.; Ji, S.; and Coley, C. W. 2025. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. In *Forty-second International Conference on Machine Learning*.

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.

CCTE, EPA. 2019. Distributed Structure-Searchable Toxicity (DSSTox) Database. Dataset. The United States Environmental Protection Agency's Center for Computational Toxicology and Exposure.

DeBerardinis, R. J.; and Thompson, C. B. 2012. Cellular metabolism and disease: what do metabolic outliers teach us? *Cell*, 148(6): 1132–1144.

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

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

Heinonen, M.; Shen, H.; Zamboni, N.; and Rousu, J. 2012. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28(18): 2333–2341.

Hussain, M. S.; Zaki, M. J.; and Subramanian, D. 2022. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 655–665.

Hussain, M. S.; Zaki, M. J.; and Subramanian, D. 2024. Triplet Interaction Improves Graph Transformers: Accurate Molecular Graph Learning with Triplet Graph Transformers. In *International Conference on Machine Learning*.

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.

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

Liu, L.; He, D.; Fang, X.; Zhang, S.; Wang, F.; He, J.; and Wu, H. 2022. GEM-2: next generation molecular property prediction network by modeling full-range many-body interactions. *arXiv preprint arXiv:2208.05863*.

Newgard, C. B. 2017. Metabolomics and metabolic diseases: where do we stand? *Cell metabolism*, 25(1): 43–56.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; et al. 2020. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11: 565644.

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.

Rao, J.; Lin, H.; Chen, L.; Xie, J.; Zheng, S.; and Yang, Y. 2025a. Multi-modal Contrastive Learning with Negative Sampling Calibration for Phenotypic Drug Discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30752–30762.

Rao, J.; Lin, H.; Xie, J.; Wang, Z.; Zheng, S.; and Yang, Y. 2025b. Incorporating Retrieval-based Causal Learning with Information Bottlenecks for Interpretable Molecular Graph Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2398–2409.

Rao, J.; Xie, J.; Yuan, Q.; Liu, D.; Wang, Z.; Lu, Y.; Zheng, S.; and Yang, Y. 2024. A variational expectation-maximization framework for balanced multi-scale learning of protein and drug interactions. *Nature Communications*, 15(1): 4476.

Rao, J.; Xu, D.; Wei, W.; Chen, Y.; Yang, M.; and Yang, Y. 2025c. Quadruple Attention in Many-body Systems for Accurate Molecular Property Predictions. In *Forty-second International Conference on Machine Learning*.

Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; and Steinbeck, C. 2021. COCONUT online: collection of open natural products database. *Journal of Cheminformatics*, 13(1): 2.

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.

Tanaka, W.; and Arita, M. 2018. Physicochemical prediction of metabolite fragmentation in tandem mass spectrometry. *Mass Spectrometry*, 7(1): A0066–A0066.

Vander Heiden, M. G.; Cantley, L. C.; and Thompson, C. B. 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930): 1029–1033.

Wang, Y.; Chen, X.; Liu, L.; and Hassoun, S. 2025. MADGEN: Mass-Spec attends to De Novo Molecular generation. In *The Thirteenth International Conference on Learning Representations*.

Wang, Y.; Wang, T.; Li, S.; He, X.; Li, M.; Wang, Z.; Zheng, N.; Shao, B.; and Liu, T.-Y. 2024. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1): 313.

Winter, R.; Montanari, F.; Noé, F.; and Clevert, D.-A. 2019. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6): 1692–1701.

Wishart, D. S. 2019. Metabolomics for investigating physiological and pathophysiological processes. *Physiological reviews*.

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, J.; Wang, Y.; Rao, J.; Zheng, S.; and Yang, Y. 2024. Self-supervised contrastive molecular representation learning with a chemical synthesis knowledge graph. *Journal of Chemical Information and Modeling*, 64(6): 1945–1954.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34: 28877–28888.