MOL-Mamba: Enhancing Molecular Representation with Structural & Electronic Insights

Anonymous submission

Abstract

Molecular representation learning plays a crucial role in various downstream tasks, such as molecular property prediction and drug design. To accurately represent molecules, Graph Neural Networks (GNNs) and Graph Transformers (GTs) have shown potential in the realm of self-supervised pretraining. However, existing approaches often overlook the relationship between molecular structure and electronic information, as well as the internal semantic reasoning within molecules. This omission of fundamental chemical knowledge in graph semantics leads to incomplete molecular representations, missing the integration of structural and electronic data. To address these issues, we introduce MOL-Mamba, a framework that enhances molecular representation by combining structural and electronic insights. MOL-Mamba consists of an Atom & Fragment Mamba-Graph (MG) for hierarchical structural reasoning and a Mamba-Transformer (MT) fuser for integrating molecular structure and electronic correlation learning. Additionally, we propose a Structural Distribution Collaborative Training and E-semantic Fusion Training framework to further enhance molecular representation learning. Extensive experiments demonstrate that MOL-Mamba outperforms state-of-the-art baselines across eight molecular datasets. Visualization experiments also provide insights into what MOL-Mamba has learned. We will release the code after the anonymous review process.

Introduction

Molecular property prediction is a critical task in various fields, including drug discovery, material science, and chemical engineering. Accurate molecular representation is essential for predicting molecular properties and facilitating the design of novel molecules. Self-supervised pretraining methods (Li et al. 2022; Fang et al. 2022b; Luo, Shi, and Thost 2023a; Liu et al. 2022) have shown significant promise in this area, as they leverage vast amounts of unlabeled data to learn useful representations without the need for extensive human annotations. These methods can capture complex patterns and relationships within molecular structures, making them highly valuable for downstream tasks.

Early molecular representation is feature-driven, with research focusing on the computation of chemical molecular fingerprints and descriptors (Rogers and Hahn 2010; Wu et al. 2018), where molecular descriptors are numerical molecular chemical constants that can be quickly obtained by mathematical logic calculations (Moriwaki et al. 2018). These descriptors include quantitative physical, chemical, or topological features of the molecule and are essential for summarizing our understanding of molecular behavior. In recent years, data-driven molecular representations have received great popularity due to the rapid development of graph neural networks (GNNs) (Hu et al. 2020) and Transformer sequence models (Transformers) (Vaswani et al. 2017). Largely, the data-driven molecular pretraining focus on learning representations from molecular structures. Methods such as PretrainGNN (Hu et al. 2020), GROVER (Rong et al. 2020), and GEM (Fang et al. 2022a) have utilized GNNs and self-supervised learning techniques to explore 2D topologies and 3D spatial information of molecules. These approaches have achieved notable success by predicting node attributes, graph motifs, and geometric features, thereby enhancing the structural representation of molecules. However, this does not mean that researchers discard the study of features such as molecular descriptors; on the contrary, molecular representation has evolved towards a more multimodal fusion. KCL (Fang et al. 2022b) and TDCL (Luo, Shi, and Thost 2023b) facilitate the fusion of expression with molecular structure information by introducing chemical feature knowledge, thus significantly enhancing the molecular learning of the model.

Currently, the methodological study of structural and electronic feature of fusion molecules as an emerging trend still has a lot of room for development, and this paper adapts to this trend and further pushes it forward. Technically, most of the existing molecular representation methods are based on GNNs, and Graph Transformer architectures (GTs) (Ying et al. 2021). While GNNs excel in capturing local structures, they often struggle with long-range dependencies due to over-squashing. On the other hand, GTs effectively model global relationships but can lack the necessary inductive biases for graph structures. Inspired by the power of state-space models (Li et al. 2024), such as mamba (Gu and Dao 2023), a powerful successor to Transformers, for sequential causal inference. In our work, we combine the strengths of both GNNs and GTs by introducing Mamba-enhanced graph learning. This novel approach leverages Mamba's ability to retain contextual information across distant nodes, addressing the limitations of conventional GNNs. By integrating Mamba, we design a graph learning framework that sig-

nificantly improves the model's capacity to capture complex molecular structures. During the fusion of structural and electronic features, we employ a hybrid mechanism that combines Mamba with traditional attention techniques. This integration allows for a more comprehensive representation of molecular properties, effectively balancing the strengths of local and global information processing. By incorporating Mamba's advanced capabilities, our architecture enhances the model's performance in accurately predicting molecular properties, paving the way for more effective applications in drug discovery and material science.

Specifically, in this paper, we introduce MOL-Mamba (Molecular Mamba), a novel framework designed to enhance molecular representation by integrating structural and electronic insights. MOL-Mamba consists of an Atom & Fragment Mamba-Graph (MG) for hierarchical structural reasoning and a Mamba-Fusion Encoder (MF) for combining molecular structure and electronic correlation learning. Our approach also incorporates a Structural Distribution Collaborative Training and E-semantic Fusion Training framework to further refine molecular representation learning. By implementing internal semantic reasoning at the fragment and atom levels and fusing structural and electronic data, MOL-Mamba provides a more comprehensive understanding of molecular properties.

In summary, our contributions are threefold:

- We propose MOL-Mamba, a new Mamba-enhanced framework that integrates structural and electronic information for improved molecular representation learning.
- We introduce novel training strategies, including Structural Distribution Collaborative Training and E-semantic Fusion Training, to enhance the fusion of molecular data.
- Extensive experiments demonstrate that MOL-Mamba outperforms state-of-the-art baselines on multiple molecular datasets, providing superior performance and interpretability in molecular property prediction tasks.

Related Work

Molecular Representation Learning

Molecular representation learning, as the foundation for molecular property prediction, has always been a popular research area. Early feature-driven methods included studies on molecular fingerprints (Rogers and Hahn 2010; Le et al. 2020), electronic & structural descriptors (Stanton and Jurs 1990; Moriwaki et al. 2018). In recent years, data-driven molecular representation methods have gained more attention, directly leveraging molecular SMILES (Zhou et al. 2023; Shen et al. 2024), 2D topological structures (Xu et al. 2019; Liu et al. 2022), and 3D geometric conformations (Schütt et al. 2017; Klicpera, Groß, and Günnemann 2020; Li et al. 2023). It is worth noting that the current trend does not completely abandon feature-driven approaches; rather, it focuses on constructing more integrated models that fuse multiple molecular modalities (Zhou et al. 2023). KCL (Fang et al. 2022b) introduces the chemical element knowledge into the molecular structural graph for knowledge-enhanced molecule learning. TDCL (Luo, Shi, and Thost 2023a) calculates the persistent homology fingerprint (a topological feature of the molecule) as an efficient supervision for pre-training molecular structures. In order to facilitate the study of multimodal molecular representation, we work on the mining of molecular multi-view structures and their integration with molecular electronic descriptors.

Transformers & Mambas for Molecule Learning

Transformers (Vaswani et al. 2017; Luo et al. 2023) have achieved great success in sequence modeling. Graph Transformers (Ying et al. 2021; Zhou et al. 2023) allows each node to attend to all others, enabling to effectively capture long-range dependencies, avoiding over-aggregation in local neighborhoods like MPNNs (Gilmer et al. 2017; Yang et al. 2021). Recently, as a powerful successor to Transformers, Mamba (Gu and Dao 2023) has become more friendly to handling long sequences with a linear scale, and its state space model (SSM) has further enhanced sequence inference capabilities, making it a promising sequence modeling solution. Some works (Behrouz and Hashemi 2024a; Wang et al. 2024; Li et al. 2024) have combined Mamba with graphs to improve the model's nonlinear modeling ability, particularly in graph-based prediction tasks such as social networks. Motivated by these works, we attempt to introduce a new Mamba-based model namely MOL-Mamba, into molecular graph and sequence tasks, to explore a novel molecular representation method.

Methodology

Problem Formulation

Generally, a molecule contains both structural and electronic attributes. The molecular structure can be represented as a *atom-level* graph $\mathcal{G}_A = \{\mathcal{V}_A, \mathcal{E}_A, \mathcal{C}_A\}$, where $\mathcal{V}_A \in \mathbb{R}^{l \times d^a}$, $\mathcal{E}_A \in \mathbb{R}^{2 \times q}$ and $\mathcal{C}_A \in \mathbb{R}^{l \times 3}$ denote the node (atom) set, edge (bond) set and the position matrix for atoms, respectively. We can mask some edges of the graph $\mathcal{G}_A$ to generate the molecular fragment subgraphs $\mathcal{F} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots\}$. Let these fragments as a node unit, and the inter-fragment bonds as edges, the *fragment-level* graph can be represented as $\mathcal{G}_F = \{\mathcal{V}_F, \mathcal{E}_F\}$. Thus, we have two views of the molecular graphs: atom-level graph $\mathcal{G}_A$ and fragment-level graph $\mathcal{G}_F$. The molecule in electron view can be represented as $\mathcal{D}_E \in \mathbb{R}^M$ with M different electronic descriptors (e-descriptors). In the setting of self-supervised molecular representation learning, our goal is to learn graph and e-descriptor encoders $f : \{\mathcal{G}_A, \mathcal{G}_F, \mathcal{D}_E\} \mapsto \mathbb{R}^d$ which maps the input graphs and e-descriptors to a vector representation without any label information. The learned encoders can then be used for various downstream property prediction tasks through finetuning.

Bi-Level Graph Collaboration Learning

Fragment-level Graph Construction. Graph fragmentation plays a fundamental role in the quality of the learning models because it dictates the global connectivity patterns. Here, we borrow the Principal Subgraph Mining (PSM) algorithm (Kong et al. 2022) to quickly and efficiently decouple the structure of molecules, where all subgraphs have

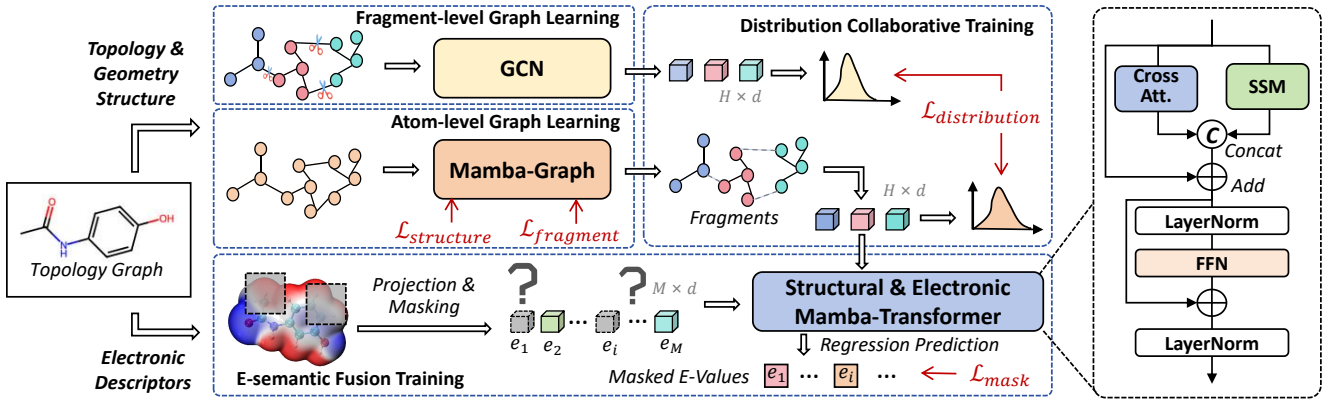

Figure 1: **The pre-training framework of our approach MOL-Mamba (Molecular Mamba).** It consists of three modules: fragment-level graph $\mathbf{GNN}_F$, atom-level structural reasoning Mamba-graph (MG), and a molecular structural & electronic Mamba-Transformer (MT) fuser. We implement two pre-training stages: molecular structure learning with the Distribution Collaborative Training, then E-semantic Fusion Training is conducted for molecular structural & electronic fusion learning.

their unique identifier ids. Subsequently, h generated molecular fragment subgraphs $\mathcal{F} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_h\}$ serve as the nodes $\mathcal{V}_F$ of fragment-graph $\mathcal{G}_F$, where $\mathcal{S}_i = \{\tilde{\mathcal{V}}^{(i)}, \tilde{\mathcal{E}}^{(i)}\}$ is a fragment subgraph with $\tilde{\mathcal{V}}^{(i)} \cap \tilde{\mathcal{V}}^{(j)} = \emptyset$, $\cup_{i=1}^{h} \tilde{\mathcal{V}}^{(i)} = \mathcal{V}_A$. An edge exists between two fragment nodes if there exists at least a bond interconnecting atoms from the fragments. Formally, $\mathcal{E}_F = \{(i,j) \mid \exists u, v, u \in \tilde{V}^{(i)}, v \in \tilde{V}^{(j)}, (u,v) \in \mathcal{E}_A\}$. After that, the fragment-graph $\mathcal{G}_F = \{\mathcal{V}_F, \mathcal{E}_F\}$ is constructed to explicitly represents the higher-order connectivity between the large components within a molecule, as shown in Figure 1 (fragment-level graph learning), supplementing molecular structure learning. We use the embedding of the identifier id of the subgraph as the initialization of the fragment node, and exploit the local learning advantage of GNNs (Kipf and Welling 2017; Xu et al. 2019) to explicitly extract high-order semantics on the structure of molecular fragments

$$\mathcal{F}_F = \text{GNN}_F(\text{LN}(\mathcal{V}_F, \mathcal{E}_F)) \in \mathbb{R}^{h \times d^f}, \tag{1}$$

where LN is LayerNorm operation (Vaswani et al. 2017), d^f is the fragment feature dimension.

Atom-level Mamba-Graph Construction. Conventional GNNs directly on the atom-graph suffers from over-squashing and poor capturing of long-range dependencies issues (Behrouz and Hashemi 2024b; Ding et al. 2024), especially for molecules with a high atomic number. To augment the model's ability to inductively correlate structural contexts within the molecular graph, and inspired by the success of Mamba (Gu and Dao 2023) in powerful and flexible sequence context modeling, we combine GNN and Mamba namely Mamba-Graph (MG) to learn structural associations within the whole molecule graph, which is more complex compared to explicit fragment structure learning, as shown in Figure 2. Specifically, we use the LN and GNN layer (Zhang, Liu, and Xie 2020) to initialize the node embedding of atom-graph $\mathcal{G}_A$ with its position information $\mathcal{C}_A$, then we propose a novel **graph node sorting strategy**, by considering molecular fragment structure and the node degree information (reflecting the importance of the node), so that the data structure is adapted to exploit Mamba's context-aware reasoning:

$$\mathcal{F}_A^G = \text{GNN}_A(\text{LN}(\mathcal{V}_A, \mathcal{E}_A, \mathcal{C}_A)) \in \mathbb{R}^{h \times d^a}; \tag{2}$$

$$\mathcal{F}_A^S = \text{SORT}_D(\text{SORT}_F(\mathcal{F}_A^G)), \tag{3}$$

where d^a is the atom feature dimension, SORT_F and SORT_D are fragment-based and node-degree-based sorts, respectively. Although Mamba (Gu and Dao 2023) employs selective SSM (State Space Model) to overcome the time-invariant nature of the system like Transformer, we add **positional encoding** to the atom sequences fed into the Mamba system, to enhance the atomic position-aware capability of the MG module, as follows:

$$\mathcal{F}_A^P = \mathcal{F}_A^S \| \mathcal{P}_F \| \mathcal{P}_D, \tag{4}$$

where $\|$ denotes concatenation, $\mathcal{P}_F$ encodes the fragment label for each atom, reflecting the fragment to which each atom belongs. $\mathcal{P}_D$ encodes the intra-fragment positional information for atoms within the same fragment.

In order to better delve the structure information of graph, correlate and synchronize the graph and Mamba, we design a new **GSSM (GraphSSM) mechanism** as shown in Figure 2, it allows the model to adaptively select relevant information from the graph context, with a Graph Information FeedForward workflow. Here, we feed the structural information of the graph (Adjacency matrix $\mathcal{A}_G \in \mathbb{R}^{l \times l}$ and distance matrix $\mathcal{D}_G \in \mathbb{R}^{l \times l}$) into the GSSM of Mamba. Specially, the GSSM can be achieved by remaking the SSM parameters ($\boldsymbol{A}$, $\boldsymbol{B}$ and Δ) as the new functions of the input data sequence, we express the process in Algorithm 1. The input of the Mamba block is $x = \mathcal{F}_A^P \in \mathbb{R}^{h \times d^a}$, as x passes through it, we perform the FragPool (MaxPool within each fragment) to obtain the fragment-level graph embedding:

$$\mathcal{F}_A^M = \text{FragPool}(\text{Mamba}(\mathcal{F}_A^P)) \in \mathbb{R}^{H \times d}. \tag{5}$$

In addition, two fragment-related structural losses are used for self-supervised constrained model training to facilitate

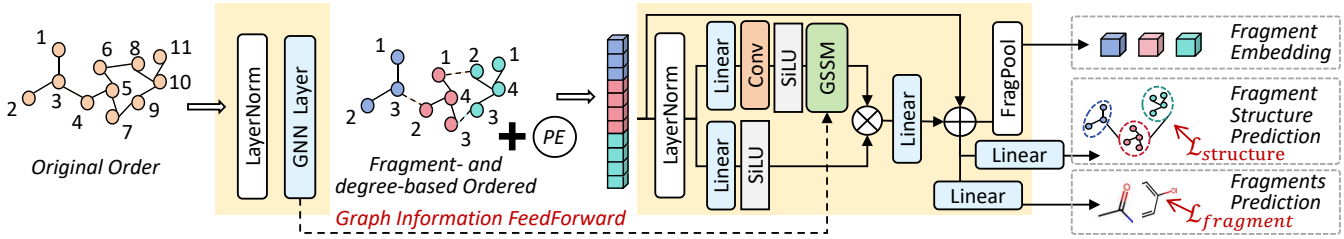

Figure 2: **Illustration of the Mamba-Graph (MG) module.** It consists of the GNN layer and Mamba block. A novel **graph node sorting strategy** to exploit Mamba's context-aware reasoning. A new **GSSM mechanism** is designed to adaptively select relevant information from the graph context. Two fragment-related loss constraints $\mathcal{L}_{structure}$ and $\mathcal{L}_{fragment}$ facilitate molecular structure learning of the MG module.

Algorithm 1: Mamba block with Graph SSM

Input: $x \in \mathbb{R}^{b\times l\times d}$
Output: $y \in \mathbb{R}^{b\times l\times d}$
Feedforward graph structure information from GNN
$\mathcal{A}_G, \mathcal{D}_G \in \mathbb{R}^{l\times l} \leftarrow$ Adjacency and distance matrices
$x, z \in \mathbb{R}^{b\times l\times d^h} \leftarrow \text{Linear}(x), \text{Linear}(x)$
$x \in \mathbb{R}^{b\times l\times d^h} \leftarrow \text{SiLU}(\text{Conv1d}(x))$
Initialize State Transition Matrix $\boldsymbol{A}$
$\boldsymbol{A} \in \mathbb{R}^{d\times n} \leftarrow$ Parameter
Initialize Control Matrix $\boldsymbol{B}$ and $\boldsymbol{C}$
$\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{b\times l\times n} \leftarrow \text{Linear}(x), \text{Linear}(x)$
Parameterize discretization step size Δ
$\Delta \in \mathbb{R}^{b\times l\times d} \leftarrow \text{Softplus}(\text{Linear}(x))$
Integrate Δ, $\mathcal{A}_G$ and $\mathcal{D}_G$, with Hadamard product $\odot$
$\Delta' \leftarrow (\Delta \odot \mathcal{A}_G \odot \mathcal{D}_G)$
Discretize continuous parameters A and B:
$\overline{\boldsymbol{A}} \in \mathbb{R}^{b\times l\times d\times n} \leftarrow \exp(\sum_{i,j} \Delta'_{ij} \boldsymbol{A}_{ij})$
$\overline{\boldsymbol{B}} \in \mathbb{R}^{b\times l\times d\times n} \leftarrow \sum_{i,j,k} \Delta'_{ij} B_{jk} x_{ki}$
Perform SSM with the graph information
$y \in \mathbb{R}^{b\times l\times d^h} \leftarrow \text{SSM}(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C})(x)$
$y \in \mathbb{R}^{b\times l\times d^h} \leftarrow y \odot \text{SiLU}(z)$
$y \in \mathbb{R}^{b\times l\times d} \leftarrow \text{Linear}(y) + x$
return y

structural inference in the MG module. Thanks to the optimized fragmentation procedure (Kong et al. 2022) that we use, we can enumerate all the trunks of the fragment-graph and use the output of Mamba block (full molecular graph information) to predict it and all subgraph's ids, as two multi-label prediction tasks, with the loss of $\mathcal{L}_{structure}$ ($\mathcal{L}_s$) and $\mathcal{L}_{fragment}$ ($\mathcal{L}_f$), respectively.

Structure Collaborative Training. To train both fragment-level and atom-level graphs, we maximize the consistency between local structural information of molecular fragments and the global structural information. Specifically, we conduct the Softmax function on the $\mathcal{F}_F$ from fragment-graph GNN_F as a pseudo-label for the fragment distribution. Also, based on the output fragment embedding $\mathcal{F}_A^M$ of MG module, we calculate the predicted labels. The objective function $\mathcal{L}_{distribution}$ ($\mathcal{L}_d$) is

$$\hat{\mathcal{F}}_F = \text{SoftMax}(\tau \cdot \mathcal{F}_F), \hat{\mathcal{F}}_A = \text{SoftMax}(\tau \cdot \mathcal{F}_A^M); \quad (6)$$

$$\mathcal{L}_d = -(\mathbb{E}_{\hat{f}_u^F, u\in\mathcal{V}_F} \log(\hat{f}_u^A) + \mathbb{E}_{\hat{f}_u^A, u\in\mathcal{V}_A} \log(\hat{f}_u^F)), \quad (7)$$

where τ is a temperature coefficient used to control the smoothness of the soft labels. $\mathcal{L}_d$ is the cross-entropy loss designed to encourage mutual supervision between the GNN_F and MG modules.

E-semantic Driven Structure & Electron Fusing

Molecular Electronic View Expression. We statistics the descriptors that represent the molecules from the electronic view in Table 1. Then the molecular electronic view expression is $\mathcal{D}_E = \mathcal{D}_S \| \mathcal{D}_M \| \mathcal{D}_Q \| \mathcal{D}_C \in \mathbb{R}^{M\times 2}$. To integrate molecular structure and electronic information, we design an E-semantic masked training strategy, *i.e.*, we randomly mask $\alpha\%$ of the input sequence $\mathcal{D}_E$ with the mask matrix $\mathcal{M}$, then the MLPs serves as the E-encoder to convert the masked descriptor tokens $\mathcal{D}_E^M$ into a sequence of tokens $\mathcal{S}_E^M \in \mathbb{R}^{M\times d}$ for being further processed by Mamba-Transformer model.

Type of Descriptors	Notation	Dimension	Number
E-state	$\mathcal{D}_S$	2	25
Molecular Property	$\mathcal{D}_M$	2	55
Quantum Chemical	$\mathcal{D}_Q$	2	7
Charge	$\mathcal{D}_C$	2	25

Table 1: The statistics of Electrochemical Descriptors.

Unified Mamba-Transformer Fuser. For each molecule, we combine its structural and electronic representations. We adopt the Mamba-Transformer (MT) backbone to integrate these two features, the MT module is shown in Figure 1 (right). Here, the MT module is a state space augmented transformer variant to enhance modeling of long sequences (concatenation of structure and electronic embeddings). Based on the masked electronic feature $\mathcal{S}_E^M$, we introduce a self-supervised mask prediction task, to enhance the interaction of structural and electronic features and to remove the effect of redundant electronic descriptors for

downstream task prediction. The processes are

$$\mathcal{U} = \text{Transformer}(\mathcal{S}_E^M \| \mathcal{F}_A^M) \in \mathbb{R}^{(M+H)\times d}; \quad (8)$$

$$\hat{\mathcal{S}}_E^M = \text{MLP}(\mathcal{U}[1:M,:]) \in \mathbb{R}^{M\times d}; \quad (9)$$

$$\mathcal{L}_{mask} = \frac{1}{\sum_{i=1}^{M} \mathcal{M}_i} \sum_{i=1}^{M} \mathcal{M}_i \Big((\mathcal{S}_E^M)_i - (\hat{\mathcal{S}}_E^M)_i \Big)^2. \quad (10)$$

Optimization Objective

Pre-training. In our pre-training phase, we define four loss functions, each with specific optimization objectives. The overall objective is to minimize a weighted sum of these losses. The four losses are: structural distribution collaborative training loss $\mathcal{L}_d$ of GNN$_F$ and MG module, two fragment-related structural losses $\mathcal{L}_s$ and $\mathcal{L}_f$ for MG module, masked E-semantic fusion training loss $\mathcal{L}_{mask}$. The overall optimization objective combines these four loss functions with specific weights:

$$\mathcal{L}_{\text{total}} = \lambda_d \mathcal{L}_d + \lambda_s \mathcal{L}_s + \lambda_f \mathcal{L}_f + \lambda_{mask} \mathcal{L}_{mask} \quad (11)$$

where λ_d, λ_s, λ_f, and λ_{mask} are the weights for each loss component, respectively.

Downstream Inference. After pre-training, the learned encoders (GNN$_F$, MG module, E-encoder(MLPs) and MT fuser) are fine-tuned on specific downstream property prediction tasks. Note that the output of the first three encoders is concatenated and injected into MT fuser. Subsequently, a two-layer MLPs is attached to the MT module as a prediction head for downstream tasks. The fine-tuning process ensures that the model adapts to the specific requirements of each downstream task, leveraging the rich structural and electronic information captured during pre-training.

Experiments

In this section, we conduct comprehensive experiments to demonstrate the effcacy of our proposed method. The experiments are designed to analyze the method by addressing the following key questions:

- **Q1:** How does MOL-Mamba perform compared with state-of-the-art methods for molecular property prediction?
- **Q2:** How do the fragment-graph GNN$_F$, the atom-level MG module and the MT fuser affect MOL-Mamba?
- **Q3:** Do $\mathcal{L}_{distribution}(\mathcal{L}_d), \mathcal{L}_{structure}(\mathcal{L}_s), \mathcal{L}_{fragment}(\mathcal{L}_f)$ and $\mathcal{L}_{mask}$ provide useful supervision for molecular structure and electronic learning during pre-training?
- **Q4:** How useful are graph node sorting strategy, position encoding and Mamba block with GSSM in MG Module?
- **Q5:** Quantitative mechanistic interpretation of the model.

Experiment Settings

Datasets. We use the recently popular GEOM (Axelrod and Gomez-Bombarelli 2022) that contains 50k qualified molecules, for molecular pretraining, followed by (Liu et al. 2022; Wang et al. 2023). For downstream tasks, we conduct experiments on 11 benchmark datasets from the MoleculeNet (Wu et al. 2018), they involve physical chemistry, biophysics, physiology and quantum mechanics. Based on task type, 7 classification benchmarks (BBBP, Tox21, ClinTox, HIV, BACE, SIDER, MUV) and 4 regression benchmarks (FreeSolv, ESOL, Lipo, QM9) are included. Each dataset uses the recommended splitting method to divide data into training/validation/test sets with a ratio of 8:1:1.

Baselines. We compare the proposed MOL-Mamba against multiple baselines, including both supervised and self-supervised/pre-training methods. **(1) Supervised GNN methods** include SchNet (Schütt et al. 2017), GIN (Xu et al. 2019), AttentiveFP (Xiong et al. 2019), and DMPNN (Yang et al. 2021). SchNet models molecular quantum interactions using continuous-filter convolutional neural networks. GIN enhances the ability to learn complex graph structures through graph isomorphism and learnable aggregation functions. DMPNN and AttentiveFP are two variants of message passing neural networks (Gilmer et al. 2017). **(2) Pre-training methods** include PretrainGNN (Hu et al. 2020), GROVER (Rong et al. 2020), GEM (Fang et al. 2022a), GraphMVP (Liu et al. 2022), MolCLR (Wang et al. 2022), Uni-Mol (Zhou et al. 2023), and MOLEBLEND (Yu et al. 2024). *PretrainGNN* introduces node-level self-supervised methods, such as context prediction and attribute masking, for pre-training GNNs. *GROVER* proposes a self-supervised graph transformer method for learning molecules through graph-level motif prediction. *GEM* employs a geometry-enhanced approach to capture molecular 3D spatial knowledge, including bond lengths, bond angles, and atomic distances. *GraphMVP* improves molecular structural representation via contrastive pretraining of both 2D topology and 3D molecular structures. *MolCLR* aligns substructures of similarly structured molecules to enhance molecular representation. *Uni-Mol* and *MOLEBLEND* integrate 1D sequence, 2D topology, and 3D information to improve model performance and adaptability for various molecular tasks.

Implementation Details. We follow the traditional metrics ROC-AUC↑ and RMSE/MAE↓ (Li et al. 2022) for the molecular property prediction classification and regression tasks respectively. And the performance metrics reported here are the averaged results of 10-fold cross-validation and the standard deviations. *In the implementation*, molecular descriptors are calculated by the ChemDes package (Dong et al. 2015). We adopt 6-layer GIN (Xu et al. 2019) as the fragment-graph GNN$_F$, and set a 6-layer SchNet (Schütt et al. 2017) as the atom-graph GNN$_A$ of the MG module, with 2-layer Mamba blocks. *For pre-training*, we set the temperature coefficient in Eq. 6 $\tau = 0.5$, and we set the mask ratio as $\alpha = 10$ (%) for mask matrix $\mathcal{M}$ in Eq. 10. Based on the order of magnitude of each loss, we set different loss weights as follows, $\mathcal{L}_d = \mathcal{L}_s = \mathcal{L}_{mask} = 0.1, \mathcal{L}_f = 20.0$, respectively. *For pre-training and fine-tuning*, we employ the AdamW optimizer, the learning rate is set to 0.0001, and the batch size is 64, and the training is conducted 100 epochs, with the early stopping on the validation set. We develop all codes on a single NVIDIA RTX A5000 GPU.

	Method	Venue	BBBP 2,039	Tox21 7,831	ClinTox 1,478	HIV 41,127	BACE 1,513	SIDER 1,427	MUV 93,087
supervised	SchNet	NIPS'17	**84.8 (±2.2)**	77.2 (±2.3)	71.5 (±3.7)	70.2 (±3.4)	76.6 (±1.1)	53.9 (±3.7)	71.3 (±3.0)
	GIN	ICLR'19	65.8 (±4.5)	74.0 (±0.8)	58.0 (±4.4)	75.3 (±1.9)	70.1 (±5.4)	57.3 (±1.6)	71.8 (±2.5)
	AttentiveFP	JMC'19	64.3 (±1.8)	**76.1 (±0.5)**	84.7 (±0.3)	**75.7 (±1.4)**	78.4 (±2.2)	60.6 (±3.2)	**76.6 (±1.5)**
	DMPNN	NIPS'21	71.2 (±3.8)	68.9 (±1.3)	**90.5 (±5.3)**	75.0 (±2.1)	**85.3 (±5.3)**	**63.2 (±2.3)**	76.2 (±2.8)
pre-training	PretrainGNN	ICLR'20	68.7 (±1.3)	78.1 (±0.6)	72.6 (±1.5)	79.9 (±0.7)	84.5 (±0.7)	62.7 (±0.8)	81.3 (±2.1)
	GROVER	NIPS'20	69.5 (±0.1)	73.5 (±0.1)	76.2 (±3.7)	68.2 (±1.1)	81.0 (±1.4)	65.4 (±0.1)	67.3 (±1.8)
	GEM	NMI'22	72.4 (±0.4)	78.1 (±0.1)	90.1 (±1.3)	80.6 (±0.9)	85.6 (±1.1)	67.2 (±0.4)	81.7 (±0.5)
	GraphMVP	ICLR'22	72.4 (±1.6)	74.4 (±0.2)	77.5 (±4.2)	77.0 (±1.2)	81.2 (±0.9)	63.9 (±1.2)	75.0 (±1.0)
	MolCLR	NMI'22	**73.6 (±0.5)**	**79.8 (±0.7)**	**93.2 (±1.7)**	80.6 (±1.1)	**89.0 (±0.3)**	**68.0 (±1.1)**	**88.6 (±2.2)**
	Uni-Mol	ICLR'23	72.9 (±0.6)	79.6 (±0.5)	91.9 (±1.8)	**80.8 (±0.3)**	85.7 (±0.2)	65.9 (±1.3)	82.1 (±1.3)
	MOLEBLEND	ICLR'24	73.0 (±0.8)	77.8 (±0.8)	87.6 (±0.7)	79.0 (±0.8)	83.7 (±1.4)	64.9 (±0.3)	77.2 (±2.3)
	MOL-Mamba (Ours)		**75.0 (±0.2)**	**81.3 (±0.4)**	**92.7 (±1.1)**	**81.6 (±0.5)**	**86.4 (±0.3)**	**68.3 (±0.9)**	**89.0 (±1.2)**

Table 2: Results of state-of-arts on seven classification benchmarks. Mean and standard deviation of test ROC-AUC↑ (%) on each benchmark are reported. Best performing supervised and self-supervised/pre-training methods for each benchmark are marked as **bold**. Second best self-supervised methods are marked as **bold**.

Method	FreeSolv 642	ESOL 1,128	Lipo 4,200	QM9 130,829
SchNet	3.22 (±0.76)	1.05 (±0.06)	0.91 (±0.10)	0.081
GIN	2.76 (±0.18)	1.45 (±0.02)	0.85 (±0.07)	0.009
AttentiveFP	**2.07 (±0.18)**	**0.88 (±0.03)**	0.72 (±0.00)	**0.008**
DMPNN	2.18 (±0.91)	0.98 (±0.26)	**0.65 (±0.05)**	0.008
PretrainGNN	2.76 (±0.00)	1.10 (±0.01)	0.74 (±0.00)	0.009
GROVER	2.27 (±0.05)	0.90 (±0.02)	0.82 (±0.01)	0.010
GEM	1.88 (±0.09)	0.80 (±0.03)	0.66 (±0.01)	**0.007**
GraphMVP	-	1.03 (±0.03)	0.68 (±0.01)	-
MolCLR	2.20 (±0.20)	1.11 (±0.01)	0.65 (±0.08)	-
Uni-Mol	**1.62 (±0.04)**	**0.79 (±0.03)**	**0.60 (±0.01)**	**0.005**
MOL-Mamba	**1.02 (±0.02)**	**0.63 (±0.01)**	**0.53 (±0.01)**	**0.007**

Table 3: Results of state-of-arts on four regression benchmarks. Mean and standard deviation of test RMSE↓ (for FreeSolv, ESOL, Lipo) or MAE↓ (for QM9) are reported. Since the standard deviation of all results on QM9 is close to 0, we do not show them.

Performance Evaluation

Overall Comparision (Q1). Table 2 and 3 present the experimental results of MOL-Mamba compared to competitive baselines, with the best results highlighted in bold. Our MOL-Mamba framework demonstrates superior performance on 8 out of 11 classification and regression benchmarks. For instance, it achieves a 75.0% ROC-AUC with a standard deviation of 0.2 on the BBBP dataset, and the lowest mean MAE of 0.63 on the ESOL dataset. This underscores MOL-Mamba's strength as a robust pre-training framework that is both easy to implement and highly adaptable across various molecular domain tasks.

Moreover, MOL-Mamba consistently outperforms the leading supervised baselines, showcasing its ability to integrate complex molecular information effectively. In contrast to DMPNN (Yang et al. 2021), which relies on intricate molecular conformation optimization and domain-specific features, MOL-Mamba surpasses it by 1.2% on the ClinTox dataset. This highlights MOL-Mamba's capability to enhance molecular property predictions without extensive reliance on domain-specific annotations, making it a versatile tool for advancing research in molecular science. Its comprehensive integration of structural and electronic insights provides a significant edge in predictive accuracy, further establishing its value in diverse applications.

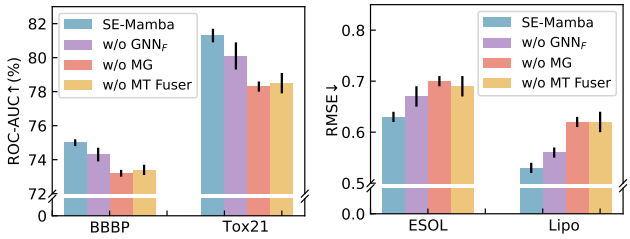

Figure 3: MOL-Mamba with different module settings on classification benchmarks BBBP and Tox21, regression benchmarks ESOL and Lipo, (lower is better for regression).

$\mathcal{L}_d$	$\mathcal{L}_s$	$\mathcal{L}_f$	$\mathcal{L}_{mask}$	BBBP	Tox21	ESOL	Lipo
-	-	-	-	72.5 (±0.4)	78.6 (±0.6)	0.70 (±0.03)	0.63 (±0.03)
✓	-	-	-	73.4 (±0.3)	79.1 (±0.5)	0.68 (±0.02)	0.59 (±0.02)
✓	✓	-	-	73.8 (±0.3)	80.5 (±0.5)	0.66 (±0.02)	0.58 (±0.02)
✓	✓	✓	-	74.2 (±0.3)	80.9 (±0.4)	0.65 (±0.02)	0.56 (±0.02)
✓	✓	✓	✓	**75.0 (±0.2)**	**81.3 (±0.4)**	**0.63 (±0.01)**	**0.53 (±0.01)**

Table 4: Mol-Mamba with different pre-training loss settings on classification benchmarks BBBP and Tox21, regression benchmarks ESOL and Lipo, (lower is better for regression).

Model	BBBP	ToX21	ESOL	Lipo
MG	**75.0 (±0.2)**	**81.3 (±0.4)**	**0.63 (±0.01)**	**0.53 (±0.01)**
w/o SORT	73.2 (±0.4)	79.5 (±0.5)	0.66 (±0.03)	0.56 (±0.03)
w/o PE	74.0 (±0.3)	80.1 (±0.5)	0.65 (±0.02)	0.55 (±0.03)
w/o GSSM	73.0 (±0.3)	78.9 (±0.6)	0.67 (±0.03)	0.57 (±0.03)

Table 5: MG module with different settings. "SORT" denotes the graph node sorting strategy, "PE" denotes positional encoding $\mathcal{P}_F$ and $\mathcal{P}_D$, "GSSM" refers to GraphSSM mechanism in the Mamba block.

Main Module Analysis (Q2). Figure 3 illustrates the critical role of MOL-Mamba's main modules, highlighting their contributions to model performance. The absence of the GNN$_F$ module results in noticeable performance drops across all datasets, underscoring its importance in capturing fragment-level structural information. Similarly, the removal of the Mamba-Graph (MG) module leads to a decline in accuracy, particularly in classification tasks like Tox21, emphasizing its role in enhancing atom-level structural reason-

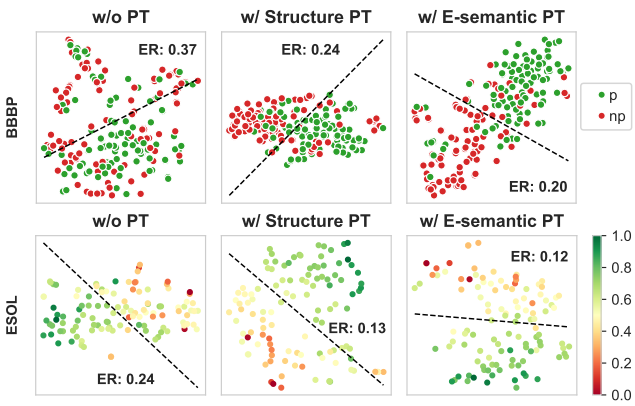

Figure 4: T-SNE visualization of feature separation of our pre-training (PT) methods on BBBP and ESOL. The "Structure PT" refers to pretraining the model with structural losses ($\mathcal{L}_d, \mathcal{L}_s, \mathcal{L}_f$), the "E-semantic PT" pretrains the model with e-masked loss $\mathcal{L}_{mask}$. "ER" is Error Rate.

ing. The MT Fuser module is pivotal for integrating molecular structure and electronic data, as evidenced by the marked decrease in both classification and regression benchmarks when it is omitted. This module's ability to fuse diverse types of molecular information is crucial for achieving comprehensive and accurate representations. The results show that each component of MOL-Mamba contributes uniquely to its robust performance, with the main modules collectively enhancing the model's capacity to predict molecular properties effectively.

Self-supervised Loss Analysis (Q3). Table 4 highlights the impact of pre-training loss settings on MOL-Mamba's performance. Without any constraints, the model underperforms, emphasizing the need for pre-training strategies. Introducing $\mathcal{L}_d$ improves results by capturing key structural features. Adding $\mathcal{L}_s$ further enhances performance, especially on Tox21, by boosting semantic-level reasoning. Incorporating $\mathcal{L}_f$ provides further gains in BBBP and Lipo. The full model, integrating all loss components including $\mathcal{L}_{\text{mask}}$, performs the best, effectively combining structural and electronic data for accurate molecular predictions.

Mamba-Graph Module Analysis (Q4). From Table 5, the full MG configuration achieves optimal results across all benchmarks, underscoring its comprehensive design. Omitting the graph node sorting strategy "SORT") results in significant performance declines, particularly on the BBBP and Tox21 datasets, highlighting its essential role in node representation. Similarly, removing positional encoding ("PE") reduces accuracy, indicating its importance for maintaining spatial information. The GraphSSM mechanism ("GSSM") is crucial, as its absence causes the largest drop, especially in regression tasks like ESOL and Lipo. These findings demonstrate that each component is vital for the MG module's effectiveness in accurately predicting molecular properties.

Feature Separation Analysis (Q5). Figure 4 presents a t-SNE visualization illustrating the feature separation

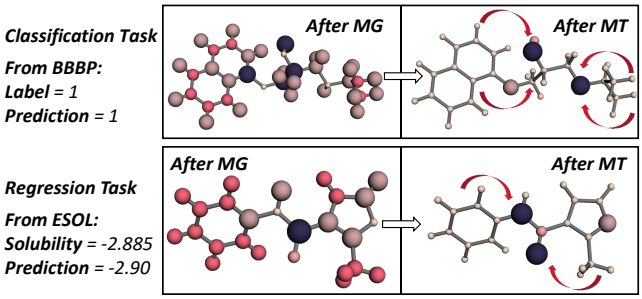

Figure 5: Exemplary explanations for using MG and MT modules in molecular prediction tasks. The radius scales linearly with the node-feature sum. It highlights key structural and electronic adjustments.

achieved by different pre-training (PT) methods on the BBBP and ESOL benchmarks. The error rate (ER), defined as the proportion of incorrectly classified points, is noted for each scenario. Without PT, the ER is highest at 0.37 for BBBP and 0.24 for ESOL, indicating poor separation between positive (p) and negative (np) samples. Introducing "Structure PT" significantly reduces the ER to 0.24 for BBBP and 0.13 for ESOL, demonstrating improved clustering due to enhanced structural understanding. The "E-semantic PT" further refines feature separation, achieving the lowest ERs of 0.20 and 0.12, respectively. This highlights the effectiveness of incorporating e-masked loss $\mathcal{L}_{\text{mask}}$ in capturing complex semantic nuances, leading to more distinct and accurate classification boundaries.

Feature Visualization Analysis (Q5). Visualization of feature weights provides insights into the relative importance of each input feature in the model's decision-making process. In Figure 5, the feature transformations through the MG and MT modules are illustrated. For the classification task from BBBP, the MG module captures essential structural features, aligning predictions accurately with labels. The MT module further adjusts features by focusing on key electronic interactions, enhancing decision-making precision. In the regression task from ESOL, the MG module effectively models solubility characteristics, while the MT module refines these features, leading to highly accurate solubility predictions. This demonstrates the model's ability to prioritize relevant molecular areas, such as reactive centers and stereochemistry. These transformations ensures that predictions are based on meaningful molecular patterns, enhancing the model's interpretability and reliability.

Conclusion

This paper introduces the MOL-Mamba framework, which significantly enhances molecular representation learning by integrating structural and electronic insights. Comprising hierarchical structural reasoning and a fusion encoder, MOL-Mamba employs innovative training strategies to achieve accurate molecular property predictions. Extensive experiments demonstrate its superior performance over state-of-the-art methods, offering a robust predictive tool for applications in drug discovery and materials science.

References

Axelrod, S.; and Gomez-Bombarelli, R. 2022. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1): 185. 5

Behrouz, A.; and Hashemi, F. 2024a. Graph Mamba: Towards Learning on Graphs with State Space Models. *CoRR*, abs/2402.08678. 2

Behrouz, A.; and Hashemi, F. 2024b. Graph mamba: Towards learning on graphs with state space models. *arXiv preprint arXiv:2402.08678*. 3

Ding, Y.; Orvieto, A.; He, B.; and Hofmann, T. 2024. Recurrent Distance Filtering for Graph Representation Learning. In *Forty-first International Conference on Machine Learning*. 3

Dong, J.; Cao, D.; Miao, H.; Liu, S.; Deng, B.; Yun, Y.; Wang, N.; Lu, A.; Zeng, W.; and Chen, A. F. 2015. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminformatics*, 7: 60:1–60:10. 5

Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; and Wang, H. 2022a. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2): 127–134. 1, 5

Fang, Y.; Zhang, Q.; Yang, H.; Zhuang, X.; Deng, S.; Zhang, W.; Qin, M.; Chen, Z.; Fan, X.; and Chen, H. 2022b. Molecular Contrastive Learning with Chemical Element Knowledge Graph. In *AAAI*, 3968–3976. 1, 2

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR. 2, 5

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. 1, 2, 3

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V. S.; and Leskovec, J. 2020. Strategies for Pre-training Graph Neural Networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 1, 5

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 3

Klicpera, J.; Groß, J.; and Günnemann, S. 2020. Directional Message Passing for Molecular Graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 2

Kong, X.; Huang, W.; Tan, Z.; and Liu, Y. 2022. Molecule Generation by Principal Subgraph Mining and Assembling. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. 2, 4

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. 2

Li, C.; Yao, J.; Su, J.; Liu, Z.; Zeng, X.; and Huang, C. 2023. LagNet: Deep Lagrangian Mechanics for Plug-and-Play Molecular Representation Learning. In *AAAI*, 5169–5177. 2

Li, L.; Wang, H.; Zhang, W.; and Coster, A. 2024. STG-Mamba: Spatial-Temporal Graph Learning via Selective State Space Model. *CoRR*, abs/2403.12418. 1, 2

Li, S.; Zhou, J.; Xu, T.; Dou, D.; and Xiong, H. 2022. GeomGCL: Geometric Graph Contrastive Learning for Molecular Property Prediction. In *AAAI*, 4541–4549. 1, 5

Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; and Tang, J. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 1, 2, 5

Luo, S.; Chen, T.; Xu, Y.; Zheng, S.; Liu, T.; Wang, L.; and He, D. 2023. One Transformer Can Understand Both 2D & 3D Molecular Data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. 2

Luo, Y.; Shi, L.; and Thost, V. 2023a. Improving Self-supervised Molecular Representation Learning using Persistent Homology. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 1, 2

Luo, Y.; Shi, L.; and Thost, V. 2023b. Improving Self-supervised Molecular Representation Learning using Persistent Homology. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 1

Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; and Takagi, T. 2018. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10: 1–14. 1, 2

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754. 1, 2

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 1, 5

Schütt, K.; Kindermans, P.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; and Müller, K. 2017. SchNet: A continuous-filter convolutional neural network for modeling

quantum interactions. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 991–1001. 2, 5

Shen, A.; Yuan, M.; Ma, Y.; Du, J.; and Wang, M. 2024. Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction. *Briefings Bioinform.*, 25(4). 2

Stanton, D. T.; and Jurs, P. C. 1990. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21): 2323–2329. 2

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008. 1, 2, 3

Wang, C.; Tsepa, O.; Ma, J.; and Wang, B. 2024. Graph-Mamba: Towards Long-Range Graph Sequence Modeling with Selective State Spaces. *CoRR*, abs/2402.00789. 2

Wang, H.; Kaddour, J.; Liu, S.; Tang, J.; Lasenby, J.; and Liu, Q. 2023. Evaluating Self-Supervised Learning for Molecular Graph Embeddings. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 5

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287. 5

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530. 1, 5

Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16): 8749–8760. 5

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2, 3, 5

Yang, S.; Li, Z.; Song, G.; and Cai, L. 2021. Deep Molecular Representation Learning via Fusing Physical and Chemical Information. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 16346–16357. 2, 5, 6

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T. 2021. Do Transformers Really Perform Badly for Graph Representation? In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 28877–28888. 1, 2

Yu, Q.; Zhang, Y.; Ni, Y.; Feng, S.; Lan, Y.; Zhou, H.; and Liu, J. 2024. Multimodal Molecular Pretraining via Modality Blending. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. 5

Zhang, S.; Liu, Y.; and Xie, L. 2020. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. *arXiv preprint arXiv:2011.07457*. 3

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. 2, 5

Checklist

1. This paper:
   (a) Includes a conceptual outline and/or pseudocode description of AI methods introduced. [YES]
   (b) Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results. [YES]
   (c) Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper. [YES]
2. Does this paper make theoretical contributions? (yes/no)
   (a) All assumptions and restrictions are stated clearly and formally. [YES]
   (b) All novel claims are stated formally (*e.g.*, in theorem statements). [YES]
   (c) Proofs of all novel claims are included. [NA]
   (d) Proof sketches or intuitions are given for complex and/or novel results. [YES]
   (e) Appropriate citations to theoretical tools used are given. [YES]
   (f) All theoretical claims are demonstrated empirically to hold. [YES]
   (g) All experimental code used to eliminate or disprove claims is included. [NA]
3. Does this paper rely on one or more datasets? (yes/no)
   (a) A motivation is given for why the experiments are conducted on the selected datasets. [YES]
   (b) All novel datasets introduced in this paper are included in a data appendix. [NA]
   (c) All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. [YES]
   (d) All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. [YES]
   (e) All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. [YES]
   (f) All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. [YES]
4. Does this paper include computational experiments? (yes/no)
   (a) Any code required for pre-processing data is included in the appendix. [NA]
   (b) All source code required for conducting and analyzing the experiments is included in a code appendix. [NO]
   (c) All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. [YES]
   (d) All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. [NO]
   (e) If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. [YES]
   (f) This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. [YES]
   (g) This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. [YES]
   (h) This paper states the number of algorithm runs used to compute each reported result. [NO]
   (i) Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. [NO]
   (j) The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). [NO]
   (k) This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. [YES]
   (l) This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. [YES]