PyrMol: A Knowledge-Structured Pyramid Graph Framework for Generalizable Molecular Property Prediction

Yajie Li[1,2], Qichang Zhao[1,2], and Jianxin Wang[1,2](✉)

[1] School of Computer Science and Engineering, Central South University, Changsha, China
[2] Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, China
jxwang@mail.csu.edu.cn

Abstract. Accurate prediction of molecular properties is essential for accelerating drug discovery. While graph neural networks (GNNs) have achieved impressive progress, most models rely on atomic graphs with limited chemical semantics, leading to suboptimal generalization across diverse biochemical tasks. Pretraining alleviates this challenge but incurs substantial costs and depends heavily on large chemical corpora. We propose PyrMol, a knowledge-structured pyramid representation learning framework that integrates multiple chemical expertise into GNNs. PyrMol constructs heterogeneous pyramid molecular graphs comprising atomic, subgraph, and molecular levels, enabling hierarchical message passing guided by functional groups, pharmacophores, and reaction-derived fragments. A multi-source knowledge fusion module adaptively aggregates heterogeneous subgraph signals, while a hierarchical contrastive strategy aligns representations across scales to enhance discrimination and consistency. Across 10 public benchmarks, PyrMol consistently outperforms 12 state-of-the-art baselines, including advanced subgraph-enhanced architectures and leading pretraining models. Moreover, incorporating pyramid graphs into existing GNNs yields universal performance gains, demonstrating the plug-and-play versatility of framework. PyrMol thus offers a principled and generalizable paradigm for chemistry-informed deep learning, highlighting the value of structured domain knowledge in molecular AI. The source code is available at https://github.com/lyj363636/PyrMol.

Keywords: Molecular Property Prediction · Heterogeneous Message Passing Neural Network · Multi-source Knowledge Fusion · Contrastive learning.

1 Introduction

The traditional drug discovery process is notoriously protracted and costly, often requiring over 15 years and exceeding $1 billion in investment for each successfully approved drug[10]. In response to these challenges, the integration of advanced artificial intelligence (AI) into drug discovery represents a significant paradigm shift, particularly in molecular property prediction as the cornerstone of pharmaceutical community[2]. AI-aided molecular property prediction facilitates the early stages of drug discovery, aiding in hit and lead compound screening, chemical structure optimization, and even novel molecular entity generation. This significantly shortens the R&D cycle, reduces costs, and minimizes human error[1]. These models capitalize on various molecular featurization methods, including sequence-based[11], structure-based[16], and image-based descriptors[28].

Notably, Graph Neural Networks (GNNs) stand out as a premier AI technique for learning directly from the graph representation of molecules[5]. Specifically, GNNs construct molecular embeddings by denoting atoms as nodes and bonds as edges, while iteratively updating node features through message propagation from neighboring nodes. A number of representative GNN models, including Graph Convolution Network(GCN)[15], Graph Isomorphism Network(GIN)[25], and Graph Attention Network(GAT)[22], possess the ability to capture different topological features within molecular graphs and have been actively explored for the molecular representation learning. In addition, recognizing the uniqueness of molecular graphs, various molecular neural networks have been specifically designed to process chemical graphs[6]. For instance, Song et al.[18] proposed Communicative Message Passing Neural Network(CMPNN), which enhanced the interactions between atoms and bonds by utilizing a communicative kernel to construct the molecular embeddings. Xiong et al.[26] designed AttentiveFP that translated the graph representation of a molecule into a continuous vector via a directed bond-based message passing approach. By applying this

message passing operation multiple times, AttentiveFP constructed higher-level bond messages to contain larger chemical substructure information. These approaches have proven to be highly effective variants, as demonstrated by their successful application and state-of-the-art results on established chemical prediction benchmarks[20,29].

However, the performance of aforementioned end-to-end models is often constrained by the limited availability of labeled molecular data. To address this bottleneck, pretraining model has emerged as a pivotal technique to mitigate the reliance on extensive annotations. Tang et al.[21] introduced a pretraining framework called MolFCL, which consisted of fragment-based contrastive learning for the pretraining stage and functional group-based prompt learning for the finetuning stage. Zang et al.[27] built Himol, which employed Multilevel Self-supervised Pretraining (MSP). MSP comprised five tasks, predicting bond connections, atomic types, bond types, atomic counts, and bond counts, to help integrate more molecular information into molecular representations from multiple perspectives. Li et al.[17] proposed KPGT, a self-supervised framework that enhanced molecular representation learning by integrating a high-capacity Line Graph Transformer with a knowledge-guided pretraining strategy. This combination synergistically captured both the precise graph structures and rich semantic knowledge of molecules. KPGT established itself as a state-of-the-art pretraining model by demonstrating superior performance across a wide range of molecular benchmarks.

Furthermore, chemists are more accustomed to understanding the causal relationship between molecular structure and properties from the perspective of chemically significant substructures, such as functional groups, rather than individual atoms or bonds[24]. Therefore, considerable research efforts[21,27,13,7,14] have focused on using functional groups or other chemical fragments to improve the predictive power. For example, Jiang et al[13] designed a pharmacophoric-constrained multi-views molecular representation graph and proposed the heterogeneous graph transformer named PharmHGT to extract vital substructures from chemical reactions. To overcome limitations like neighbors-explosion, under-reaching, and oversmoothing, Hajiabolhassan et al.[7] proposed molecular graph coarsening framework called FunQG, based on the concept of Quotient Graph. Kengkanna et al.[14] explored applicability and impact of different molecular subgraphs on model learning and interpretation, which probed and investigated the effects of different molecular subgraphs from various perspectives, including Atom, Junction Tree, Functional Group and so on. The aforementioned methods improve the molecular representation capabilities of GNN models by enhancing the original atomic graphs through substructure graphs.

Despite significant progress, critical gaps remain in the comprehensive integration of multi-source expert knowledge. Molecular properties are often jointly determined by complex factors spanning chemistry, biophysics, and other domains. However, existing methods primarily rely on simplified molecular representations, such as atomic-level or simplified graphs. This reliance on a single perspective fails to capture the diverse expert knowledge required for accurately predicting a broad range of molecular properties, resulting in models with low generalization capabilities and weak adaptability.

To overcome these limitations, we propose a knowledge-structured pyramid graph framework for generalizable molecular property prediction, PyrMol. The main contributions of our work are as follows:

- To capture molecular topological features comprehensively, we introduce PyrMol, an end-to-end heterogeneous pyramid framework, which orchestrates a novel message passing neural network across atomic, subgraph, and molecular levels, which is a plug-and-play module explicitly guided by domain knowledge to integrate multi-scale representation.
- To address the issues of poor generalization and model robustness stemming from knowledge scarcity, we propose a multi-source knowledge enhancement and fusion module, which is designed to integrate multiple streams of expertise, and balance redundancy and complementarity of knowledge flexibly, thereby improving the model's ability to handle diverse properties across different benchmarks.
- To exploit the potential of model discriminative power, we introduce a hierarchical contrastive learning strategy cross scales of molecular representation.

Extensive experiments demonstrate that the end-to-end PyrMol outperforms existing molecular graph neural networks and subgraph-enhanced methods, achieving results competitive with pretraining models. Additionally, our proposed pyramid molecular graph further improve the performance of GNNs and subgraph-enhanced methods. In summary, PyrMol provides a powerful and adaptable framework for molecular representation learning, contributing to more generalizable AI-driven drug discovery.

2 Methods

As depicted in Fig.1, PyrMol proposed by us comprises three main components. Firstly, by introducing multi-source knowledge, we construct a heterogeneous pyramid molecular graph and update multi-level molecular representations through a knowledge-guided heterogeneous message passing neural network. Secondly, we deploy a multi-source knowledge enhancement and fusion module to explore interactions and integration among diverse information. Finally, we incorporate a hierarchical contrastive learning strategy to enhance the discrimination capability of model on multi-level molecular representations. A fully connected network is used to generate prediction results.

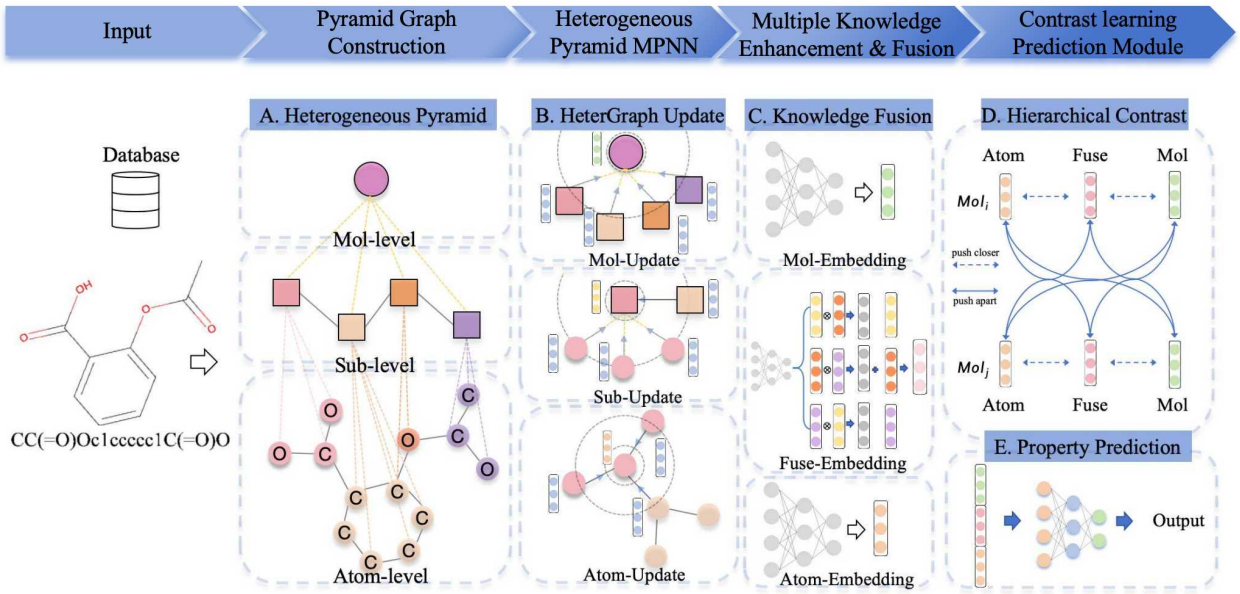

Fig. 1. Illustration of Knowledge-Structured Pyramid Graph Framework(PyrMol). PyrMol contains five modules: A. Heterogeneous Pyramid; B. HeterGraph Update; C. Knowledge Fusion; D. Hierarchical Contrast; E. Property Prediction. For a molecular, firstly, the heterogeneous pyramid molecular graph is formalized for Atom-level, Sub-leve, and Mol-level in A module. Then, heterogeneous node updates are performed in B module. Thirly, enhancement and fusion of substructure graphs guided by multi-source knowledge and generate Fuse-embedding in C module, and unifying the embedding dimension of Atom-embedding and Mol-embedding. Fourthly, hierarchical contrastive learning is performed among them in D module. Finally, the features from the three levels are concatenated and fed into the downstream task E. module to predict molecular properties .

2.1 Heterogeneous Pyramid Message Passing Neural Network

Molecular Pyramid Graph. In this section, we introduce a molecular pyramid graph, a novel construction method designed to explicitly guide GNNs in learning substructure features. As shown in Fig.1, a molecular pyramid graph is a heterogeneous diagram comprising a bottom, a middle section, and a top. Atom nodes V_A and the bonds between atoms E_A form the bottom layer, representing the original topological structure of the molecular graph. Subgraph nodes V_{S_i}, edges E_{S_i}, and E_{AS_i} constitute the core, representing key substructures or critical functional groups within the molecular graph. E_{S_i} denotes that in the i-th type of knowledge, if two subgraphs share the same atoms or have atoms directly connected, then an edge exists between them. E_{AS_i} indicates which specific atoms are included in the subgraph generated based on the i-th knowledge. E_{AS_i} means which specific atoms are included in the subgraph. This type of relationship can explicitly guide feature aggregation patterns within message passing neural networks, assisting models in extracting subgraph features. The top floor encomprises a molecular node V_M and the ties between all subgraphs and the molecule node E_{S_iM}. The molecular node directly aggregate features from subgraph nodes to perform molecular feature characterization.

We refer to three knowledge from chemistry and biophysics — functional groups (denoted as F), pharmacophores(denoted as P), and chemical reaction rules (denoted as R) — to decompose molecules into a set of fragments three times. In detail, according to chemistry knowledge, we employ the functional group defined by RDKit to obtain the related substructure V_F, which converts predefined functional groups, ring types, and atom pairs into subgraph nodes. Besides, from biophysics perspective, we gain the pharmacophoric subgraphs V_P based on the ErG algorithm [19], which incorporates six types of pharmacophore information (i.e., H-bond donor, H-bond acceptor, positive, negative, hydrophobic, and aromatic). Moreover, chemists predefined chemical reactions [3] to guide the formation and cleavage of chemical bonds. Based on chemical reaction rules, we get the fragments V_R. Based on the above description, given a molecule, we build a pyramid graph $G = \{V, E\}$, in which $V = \{V_A, V_P, V_F, V_R, V_M\}$ and $E = \{E_A, E_P, E_F, E_R, E_{AP}, E_{AF}, E_{AR}, E_{PM}, E_{FM}, E_{RM}\}$. All features are computed rapidly in silico using the open-source package RDKit. Table 1 displays the node feature description of each type of node in molecular pyramid graph.

Table 1. Node features of of each type of node in molecular pyramid graph.

Node	Dimension	Feature Type
Atom	42	Atom type, atom degree, formal charge, Chiral Tag, the total number of implicit and explicit bonds connected to the atom, and hybridization type.
Pharmacophore	6	H-bond donor, H-bond acceptor, positive, negative, hydrophobic, and aromatic
FunctionalGroup	115	Predefined types of edges, rings, and functional groups
Reaction	27	BRICS reaction type features
Molecule	300	Mol2Vec[12] embedding

Message Passing Neural Networks. Given the inconsistent dimensionality of node features across different types in heterogeneous graphs, PyrMol employs Full Connect Neural Networks(FCNNs) to standardize node features to k dimensions.

$$H_i = FC_i(F_i), i = \{A, P, F, R, M\}, \tag{1}$$

where $F_i \in \mathbb{R}^{n_i \times d_i}$ is the initial feature matrix for the n_i nodes of class i and $FC(*)$ is FCNNs.

In our MPNN, PyrMol employs two graph neural networks—GraphSAGE[8] and GAT[22]—for message passing. Specifically, the heterogeneous graph formed by atom nodes V_A and subgraph nodes V_P, V_R, V_F undergo feature updates via the GraphSAGE layer, while the heterogeneous graphs composed of substructure nodes and the molecular node V_M undergo feature updates through the GAT layer.

On each step $l \in \{0, 1, ..., L\}$ of the message passing phase, the hidden state $h_i^{(l+1)}$ of the i-th node on the GraphSAGE layer is updated as follows:

$$\begin{aligned} h_{\mathcal{N}(i)}^{(l+1)} &= \text{aggregate}\left(\left\{h_j^l, \forall j \in \mathcal{N}(i)\right\}\right), \\ h_i^{(l+1)} &= \sigma\left(W \cdot \text{concat}\left(h_i^l, h_{\mathcal{N}(i)}^{(l+1)}\right)\right), \\ h_i^{(l+1)} &= \text{norm}\left(h_i^{(l+1)}\right). \end{aligned} \tag{2}$$

The hidden state of the molecular node $h_M^{(l+1)}$ on the GAT layer is updated as follows:

$$\begin{aligned} h_M^{(l+1)} &= \sum_{j \in \mathcal{N}(M)} \alpha_j W^l h_j^l, \\ \alpha_j^l &= \text{softmax}\left(e_j^l\right), \\ e_j^l &= \text{LeakyReLU}\left(\boldsymbol{a}^T\left[W h_M^l \| W h_j^l\right]\right). \end{aligned} \tag{3}$$

where α_j is the attention score between molecular node M and subgraph node j, and a^T is a learnable vector.

After L rounds of message passing, we have obtained the node features $H^L = \{H_A, H_P, H_F, H_R, h_M\}$. By performing mean pooling on $H_i, i \in \{A, P, F, R\}$ respectively, the atom feature $h_A \in \mathbb{R}^k$, functional group feature $h_F \in \mathbb{R}^k$, pharmacophore feature $h_P \in \mathbb{R}^k$, and reaction feature $h_R \in \mathbb{R}^k$ are obtained.

2.2 Multi-source Knowledge Enhancement and Fusion Module.

As shown in Fig.1, recognizing that disparate sources of expertise contribute unequally to molecular properties, we introduce a Multi-source Knowledge Enhancement and Fusion Module. This module is designed to integrate divergent feature representations from various knowledge sources, highlighting the critical common features while orchestrating their integration in a principled manner.

Taking h_F, h_P, and h_R, we calculate the similarity between pairs of them.

$$Similarity_{ij} = Norm(h_i) \cdot Norm(h_j), \tag{4}$$

where $Norm(*)$ means the normalization function, $\cdot$ is the dot product, and $i, j \in \{F, P, R\}; i \neq j$.

Then we check if the similarity $Similarity_{ij}$ is greater than threshold α. If it is, we calculate their mutual key feature vector that are essential for them.

$$k_{ij} = \frac{1}{2}(Norm(h_i) + Norm(h_j)), \tag{5}$$

where the key feature matrix K is composed by $\{k_{FP}, k_{FR}, k_{PR}\}$.

$$h_{enhance} = \sum (Softmax(Key \times W_e) \odot Key), \tag{6}$$

where $W_e \in \mathbb{R}^{k \times 1}$ is a trainable weight and $\odot$ is element-wise dot product (hadamard product). Then, we obtain the union feature vector of knowledge according to Eq.7

$$h_u = Concat(\{FC_i(k_i), i \in \{F, P, R\}\}). \tag{7}$$

Finally, through learnable weight $W_f \in \mathbb{R}^{2k \times k}$ to balance flexibly union and enhance the characteristic, a fuse-level embedding is generated as follows.

$$h_F = FC(W_f \times Concat(h_u, h_{enhance})). \tag{8}$$

2.3 Hierarchical Contrastive Learning Strategy.

Motivated by the goal of amplifying the model's discriminative ability and unlock its strong expressive potential, we develop a hierarchical contrastive learning strategy. Concretely, the Pyrmol model, as mentioned, yields three-floor embeddings: atom-level h_A, fuse-level h_F, and mol-level h_M. We calculate the similarity Sim_{ij} of pairs of them.

$$Sim_{ij} = FC_C(h_i) \cdot FC_C(h_j), i, j \in \{A, F, M\}. \tag{9}$$

During feature extraction for a set of molecules, if h_i and h_j originate from the same molecule, we use the cross-entropy loss function to optimize Sim_{ij} toward 1; otherwise, it approaches 0. That is, we close the three level representations of same molecular, and keep these away of different molecules.

As for prediction, PyrMol employs a two-layer MLP to obtain prediction scores. We define the objective function for learning as:

$$loss = loss_{task} + \lambda \cdot loss_{contrastive}, \tag{10}$$

where λ is a hyperparameter for hierarchical contrastive learning loss and $loss_{task}$ is Binary Cross-Entropy Loss for classification tasks and Mean Square Error loss for regression tasks.

3 Experiments

3.1 Baselines

To demonstrate the advantages of PyrMol, we compare it with 12 SOTA methods as baselines. The SOTA baselines include three graph neural networks (GCNs[15], GINs[25] and GATs[22]), two molecular graph neural networks (AttentiveFP [26] and CMPNN[18]), four subgraph-enhanced models Himol_w/o_pretraing[27], HimGNN[9], PharmHGT[13], and MMGX[14]), and three pretraining-based methods (HiMol[27], KPGT[17], and MolFCL[21]). Himol_w/o_pretraing[27] is a variant model of Himol without pretraining stage.

3.2 Datasets

Driven by the need to better compare and present the effectiveness of our model, we select ten benchmark molecular datasets for experiments from MolecularNet[23] including Blood-brain barrier permeability (BBBP), BACE, HIV, ClinTox, Tox21, SIDER, and Toxcast for classification tasks, and ESOL, Freesolv, and Lipophilicity for regression tasks.

Binary classification tasks. The BBBP dataset comprises 2,035 molecules annotated with binary labels indicating blood-brain barrier permeability, making it a benchmark for predicting whether compounds can cross the blood-brain barrier. The BACE dataset includes 1,513 molecules and offers both quantitative (IC_{50}) and qualitative (binary) binding measurements for inhibitors of human β-secretase 1 (BACE-1). The HIV dataset contains 41,127 compounds screened for their ability to inhibit HIV replication.

Multi-class classification tasks. The ClinTox dataset is curated from 1,468 molecules, offering a binary classification task to distinguish FDA-approved drugs from those that failed in clinical trials owing to toxicity. The Tox21 dataset, originating from the Tox21 Challenge, provides assay results for 7,821 molecules across 12 distinct targets related to drug-induced toxicity. For the SIDER dataset, it encompasses 1,379 approved drugs, with side effects annotated and organized into 27 system organ classes. We also include the ToxCast dataset, which aggregates in vitro high-throughput screening data for 8,597 compounds over 617 toxicological tasks.

Regression tasks. For predicting key physicochemical properties, we employ three benchmarks. The ESOL dataset offers aqueous solubility measurements for 1,128 compounds. FreeSolv is a curated set of 642 molecules with corresponding experimental hydration free energies. The Lipophilicity dataset includes 4,198 compounds annotated with their experimental octanol/water distribution coefficients (logD at pH 7.4).

(GCNs[15], GINs[25] and GATs[22]), two Molecular Neural Networks (AttentiveFP [26] and CMPNN[18]), four Subgraph-enhanced models Himol_wo[27], HimGNN[9], PharmHGT[13], and MMGX[14]), and three pretraining-based methods (HiMol[27], KPGT[17], and MolFCL[21])

3.3 Implementation details.

Deng et al. [4] pointed out that the public availability of benchmark datasets introduces the risk of data leakage or selection bias, where splitting strategies might be customized to favor a particular model, thus skewing the perceived generalizability. To ensure a fair and robust comparison, we have conducted 25 repeated experiments with scaffold splitting using different random seeds. All benchmark datasets have been split as training, validation, and test sets with a ratio of 0.8/0.1/0.1. We utilize the optimal hyperparameters for all baseline methods precisely as reported in their original papers. The area under the receiver operating characteristic curve (AUC) is used as the evaluation metric for classification tasks, while Root Mean Square Error (RMSE) serve as the metric for regression tasks. Our model is implemented in Pytorch and execute on a one NVIDIA Tesla a100 (40G) GPU. The node and edge features are processed by the open-source package RDKit.

3.4 Comparison with Baselines

We have compared PyrMol against 12 baselines across 10 benchmarks. To present the concluded results, the average ranking (AvgRank), measuring the rank of model performance, is employed, as shown in Table 2. Overall, PyrMol achieved the best results in the average ranking, demonstrating optimal performance on 6 out of 10 tasks and delivering competitive results on the remaining 4 tasks. Specifically, compared with traditional graph neural networks(GCNs, GINs, and GATs) and two molecular neural networks(AttentiveFP and CMPNN), PyrMol achieves large increases of AUC in classification datasets. PyrMol get AUCs of 0.949, 0.871, and 0.811 with 6.8%, 4.2%, and 4.5% improvements to these end-to-end models on the bbbp, bace, and hiv, respectively. These results confirm that explicitly incorporating expert knowledge into molecular graph

construction and guiding molecular feature learning facilitates existing graph neural network frameworks in modeling molecular chemical environments and learning molecular representations.

In contrast to subgraph augmentation models, PyrMol outperforms Himol, HimGNN, and PharmHGT in performance. This is because PyrMol integrates multiple types of prior expertise, enabling it to handle more complex molecular property prediction tasks. PyrMol exceeds multi-subgraph-enhanced model, MMGX, verifying that carefully designed heterogeneous graph message passing networks possess stronger feature extraction capabilities than graph neural networks. Furthermore, the improvements of the model compared to the SOTA pretrained model KPGT indicates that reasonably utilizing and learning from existing prior knowledge yields more effective model performance improvements than pretraining. Additional evaluation metrics are provided in Appendix.

Table 2. Performance of PyrMol and baseline methods on molecular property prediction.

Moldes	Classification(AUC↑)						Regression(RMSE↓)				Avg
	bbbp	bace	hiv	clintox	sider	tox21	toxcast	freesolv	lipophilicity	esol	Rank↑
GCNs[15]	0.874 (0.040)	0.806 (0.055)	0.765 (0.034)	0.842 (0.064)	0.588 (0.029)	0.825 (0.012)	0.701 (0.012)	2.893 (0.818)	0.803 (0.057)	1.359 (0.527)	7.1
GINs[25]	0.881 (0.040)	0.829 (0.058)	0.761 (0.044)	0.835 (0.070)	0.596 (0.029)	0.824 (0.013)	0.698 (0.015)	3.046 (0.842)	0.764 (0.041)	2.165 (1.438)	7.2
GATs[22]	0.878 (0.037)	0.813 (0.048)	0.766 (0.037)	0.851 (0.058)	0.567 (0.023)	0.814 (0.012)	0.681 (0.014)	2.916 (0.739)	0.766 (0.051)	1.508 (0.596)	7.5
AttentiveFP[26]	0.860 (0.591)	0.820 (0.037)	0.673 (0.041)	0.746 (0.111)	0.526 (0.025)	0.768 (0.016)	0.627 (0.018)	2.274 (0.367)	0.963 (0.103)	0.893 (0.119)	8.4
CMPNN[18]	0.638 (0.143)	0.632 (0.075)	0.733 (0.080)	0.433 (0.132)	0.524 (0.020)	0.519 (0.039)	0.444 (0.010)	1.784 (0.427)	**0.627** (0.027)	0.777 (0.085)	9.1
Himol_w/o Pretrain[27]	0.839 (0.061)	0.802 (0.044)	0.749 (0.052)	0.655 (0.091)	0.580 (0.030)	0.832 (0.011)	0.708 (0.014)	3.328 (1.042)	0.815 (0.057)	1.033 (0.210)	8.2
HimGNN[9]	0.882 (0.046)	0.759 (0.123)	0.566 (0.073)	0.674 (0.169)	0.518 (0.032)	0.662 (0.125)	**0.764 (0.021)**	2.582 (0.637)	0.937 (0.083)	1.535 (0.257)	9
PharmHGT[13]	0.932 (0.029)	0.834 (0.044)	0.764 (0.040)	0.866 (0.057)	**0.617** (0.030)	**0.834 (0.011)**	0.717 (0.010)	2.361 (0.621)	0.707 (0.037)	0.956 (0.200)	3.8
MMGX[14]	0.875 (0.044)	0.798 (0.119)	0.785 (0.044)	0.597 (0.119)	0.511 (0.022)	0.827 (0.012)	0.697 (0.011)	3.714 (0.885)	0.846 (0.057)	1.791 (0.419)	9.3
HiMol_w_ Pretrain[27]	0.889 (0.031)	0.822 (0.063)	0.770 (0.039)	0.639 (0.082)	0.581 (0.028)	0.721 (0.010)	0.721 (0.010)	2.695 (0.811)	0.759 (0.050)	1.040 (0.219)	6.3
KPGT[17]	0.915 (0.029)	0.862 (0.035)	0.791 (0.036)	0.854 (0.064)	**0.617** (0.021)	0.535 (0.017)	0.515 (0.016)	2.270 (0.697)	**0.626** (0.026)	0.956 (0.157)	4.5
MolFCL[21]	0.818 (0.078)	0.677 (0.053)	0.673 (0.040)	0.453 (0.098)	0.520 (0.019)	0.630 (0.036)	0.439 (0.024)	1.777 (0.494)	0.638 (0.029)	**0.769** (0.091)	8.8
PyrMol	**0.949** (0.020)	**0.871** (0.036)	**0.811** (0.041)	**0.878** (0.034)	0.615 (0.020)	**0.834** (0.020)	0.721 (0.009)	**1.697** (0.400)	0.668 (0.040)	0.865 (0.120)	**1.8**

The bold and underlined values indicate the best and second best results, respectively. ↑ means the higher the value, the better; ↓ means the lower the value, the better.

3.5 Validation of the Pyramid Molecular Graph.

To validate the effectiveness and universality of the Pyramid Molecular Graph, we embedded it into existing models for experimental verification. In this experiment, we selected three GNNs (GCNs, GINs, GATs) and two subgraph augmentation models (Himol and PharmHGT) for evaluation across six datasets (two single classification task datasets, two multiple tasks datasets, two regression task datasets). The results are provided in Table 3. 25 of 30 group results show performance improvements. Specifically, GCNs, the simplest model architecture in graph neural networks, achieve gains across all six datasets. These findings demonstrate that GCNs can extract more meaningful structural features from pyramid molecular graphs. PharmHGT and HiMol also exhibit performance enhancements over their baseline models. These findings demonstrate that our proposed pyramid molecular graph represents a more universal molecular graph enhancement approach, where multi-level explicit node message passing further guides the feature learning efficiency of graph neural networks.

Table 3. Performance of the baselines with the Pyramid Molecular Graph.

Moldes	Classification(AUC↑)				Regression(RMSE↓)	
	bbbp	bace	sider	toxcast	freesolv	lipophilicity
GATs	0.878(0.037)	0.813(0.048)	0.587(0.030)	0.691(0.012)	2.916(0.739)	**0.766(0.051)**
GATs_PMG	**0.879(0.053)**	**0.841(0.044)**	**0.588(0.037)**	**0.701(0.011)**	**2.233(0.46)**	0.806(0.037)
GINs	0.881(0.040)	0.829(0.058)	0.594(0.030)	0.689(0.015)	3.046(0.842)	**0.764(0.041)**
GINs_PMG	**0.882(0.056)**	**0.834(0.048)**	**0.614(0.031)**	**0.710(0.010)**	**2.089(0.459)**	0.792(0.040)
GCNs	0.874(0.040)	0.806(0.055)	0.566(0.025)	0.671(0.014)	2.893(0.818)	0.803(0.057)
GCNs_PMG	**0.888(0.044)**	**0.839(0.042)**	**0.611(0.030)**	**0.714(0.013)**	**2.203(0.544)**	**0.746(0.033)**
PharmHGT	**0.932(0.029)**	0.834(0.044)	**0.617(0.030)**	0.7079(0.1200)	2.361(0.621)	0.707(0.037)
PharmHGT_PMG	0.930(0.035)	**0.846(0.056)**	0.614(0.027)	**0.715(0.010)**	**2.181(0.477)**	**0.695(0.032)**
HiMol	0.839(0.061)	0.802(0.044)	0.580(0.030)	0.708(0.014)	**2.695(0.811)**	0.759(0.050)
HiMol_PMG	**0.898(0.028)**	**0.862(0.018)**	**0.602(0.030)**	**0.724(0.010)**	3.459(0.643)	**0.710(0.037)**

The bold values indicate the best results. $*_PMG$ means models employ pyramid molecular graphs as model inputs instead of atomic graphs.

Table 4. Model performance of multiple graph combination models.

Subgraphs	Classification(AUC↑)				Regression(RMSE↓)	
	bbbp	bace	sider	toxcast	freesolv	lipophilicity
A+R	0.909(0.034)	0.852(0.053)	0.553(0.024)	0.695(0.014)	2.874(0.770)	0.703(0.0258)
A+F	0.928(0.015)	**0.859(0.039)**	0.548(0.034)	0.691(0.014)	2.541(0.477)	0.666(0.043)
A+P	0.908(0.033)	0.859(0.043)	0.551(0.041)	0.699(0.007)	2.614(0.555)	0.671(0.025)
A+R+F	0.930(0.026)	0.852(0.032)	0.553(0.031)	0.690(0.018)	2.639(0.565)	0.673(0.017)
A+R+P	0.922(0.032)	0.849(0.034)	0.551(0.027)	**0.700(0.013)**	2.540(0.582)	0.683(0.027)
A+F+P	0.930(0.020)	0.852(0.038)	0.551(0.040)	0.698(0.014)	2.520(0.439)	0.680(0.038)
A+R+F+P	**0.934(0.030)**	0.850(0.033)	**0.559(0.024)**	0.695(0.024)	**2.475(0.534)**	**0.656(0.022)**

The bold values indicate the best results. A, R, F, P means atomic graphs, chemical reaction rules, functional groups, and pharmacophores, respectively.

3.6 Effectiveness Analysis of Multi-source Knowledge Fusion.

In our research, the atom graph is the base graph as it contains fundamental topology. To compare the effects on model learning performance of different subgraph-enhanced strategies, the experiment is conducted across seven modes, which represent combinations of atomic graphs and different types of prior knowledge. For example, A+R denotes the construction of a pyramid molecular graph using atomic graphs combined with chemical reaction information. The results are provided in Table 4, where contains two Binary classification tasks datasets: bbbp, bace, two Multi-class classification tasks datasets: sider, toxcast, and two Regression tasks datasets: freesolv, lipophilicity. By comparing A+R, A+F, and A+P, it can be observed that the three types of prior knowledge (R, F, P) yield superior results on different datasets, which aligns with our initial hypothesis. Given that molecular property prediction tasks capture distinct aspects of chemical characteristics, their reliance on relevant chemical information varies. Furthermore, the results indicate that A+R+F+P demonstrated the optimal performance. Therefore, the integration of R, F, and P is essential to equip the model with a comprehensive knowledge base, enabling it to adapt to the diverse requirements of various prediction tasks.

3.7 Ablation Study

Here, we conduct an ablation study to investigate the influences of the individual modules on PyrMol. The details of two different variants of PyrMol are illustrated as follows. W/o Contrast is a model that eliminates the hierarchical contrast learning strategy and optimizes the model solely using cross-entropy or MSE loss. W/o Enhance further removes the Multi-source Knowledge Enhancement Module from the W/o Contrast configuration. The results are shown in Fig.2, where contains two Binary classification tasks datasets: bbbp, bace, two Multi-class classification tasks datasets: sider, toxcast, and two Regression tasks datasets: freesolv, lipophilicity. For W/o Enhance, we observe a significant decrease in four datasets. These results demonstrate that the Multi-source Knowledge Enhancement and Fusion Module , compared to fully

connected networks, can extract molecular commonalities driven by different knowledge sources, thereby providing stronger generalization capabilities for molecular characterization. W/o Contrast performs worse than PyrMol, indicating the necessity of hierarchical contrastive learning loss, which aims to maximize consistency among features at different levels within the same molecule while minimizing consistency between different molecules.

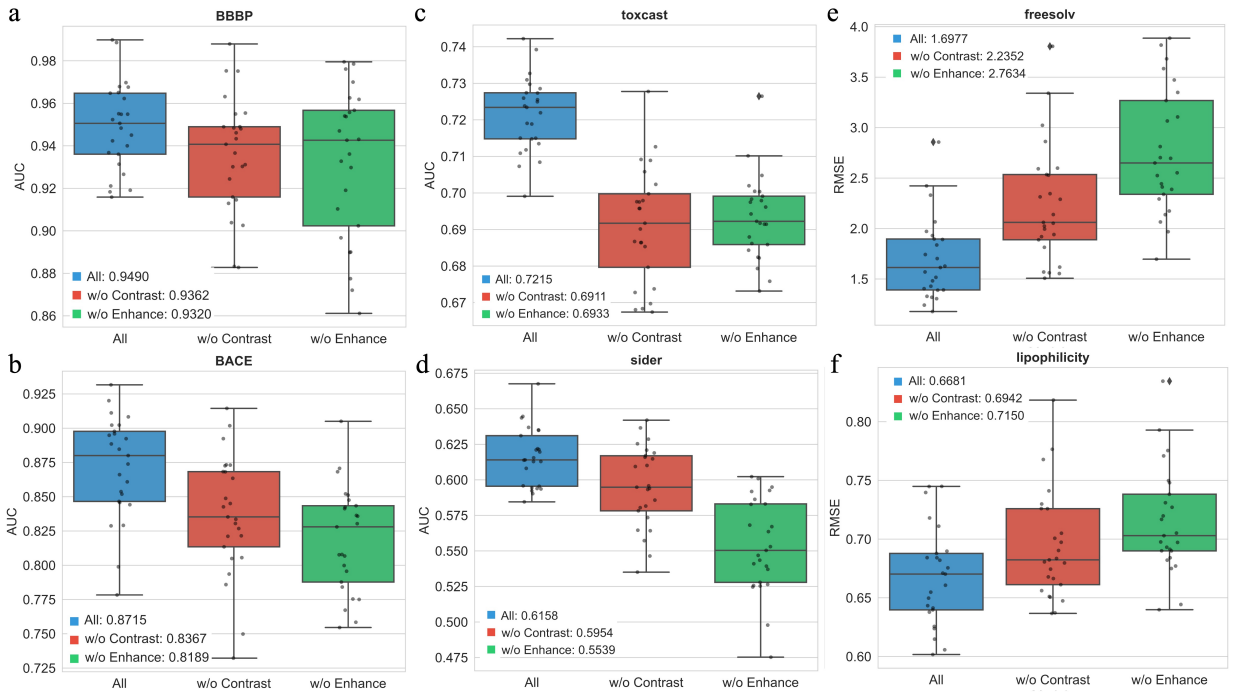

Fig. 2. The results of ablation study on six datasets. W/o Contrast is a varient that eliminates the hierarchical contrast learning strategy. W/o Enhance further removes the Multi-source Knowledge Enhancement Module from W/o Contrast.

3.8 Interpretation Results

PyrMol utilizes the GAT model to collect features for the molecular node from substructure nodes, and the relationship between atom nodes and substructure nodes is clearly defined in the pyramid graph. Therefore, PyrMol possesses the capability to explore atomic importance through attention scores. To examine the results, three examples are investigated to see how model gives attention to atom nodes. All of the following results are obtained from the models with the best performance on validation set. Examples are from BBBP dataset. As shown in Fig. 3A, the first molecule ([C@@H]3(C1=CC=CC=C1)CN2CCSC2=N3) describes two heterocycles sharing a single edge and fusing together to form a rigid planar structure. The benzene ring is attached to this fused system. The sulfur atom in the thiazole ring is less polar than the oxygen atom, resulting in stronger hydrophobicity. It typically contributes positively to BBB permeability. This should appear as red on the heatmap. The sp^2 nitrogen atom in the heterocycle acts as a strong hydrogen bond acceptor, representing the primary resistance point. This corresponds to blue on the map and contributes negatively to bbb permeability. As shown in 3B, the second molecule consists of an amino acid backbone, an aromatic side chain, a catechol hydroxyl group, and a methylene bridge. The amino acid backbone ($H_2N - CH - COOH$) is recognized by the model, which shows high correlation with BBB+. The model identifies the catechol hydroxyl group as polar, but it does not allow the negative contribution of these hydroxyl groups to outweigh the positive contribution of the entire side chain, hence the light blue color. The methylene bridge connecting the main chain and side chain ($-CH_2-$), providing hydrophobic interactions, corresponds to the red color. These results indicate that PyrMol can focus on atoms or substructures that play key roles within molecules, providing clues for molecular optimization.

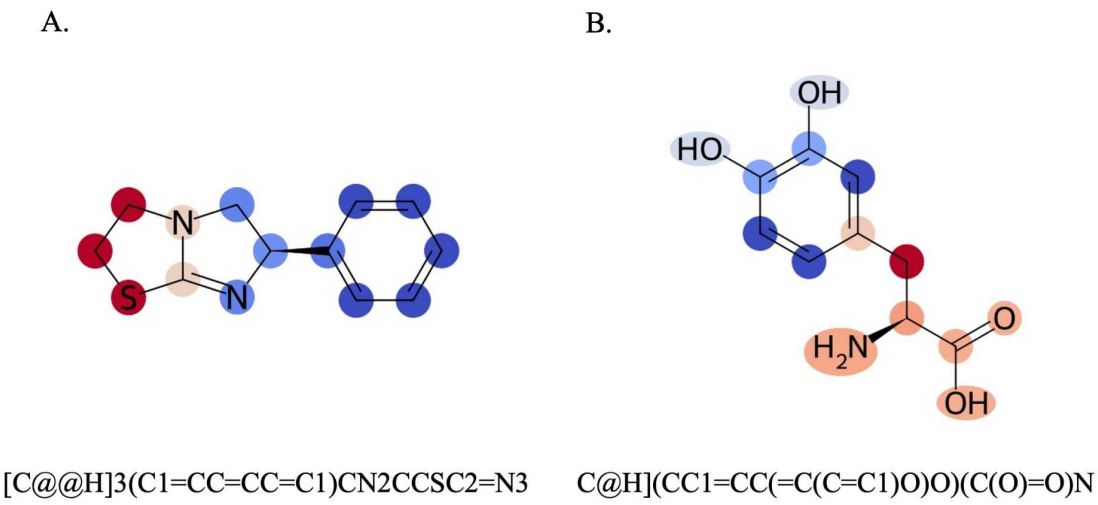

Fig. 3. Interpretability analysis of attention scores on the BBBP dataset. Dark red indicates the atom have the highest attention score, light red the next highest, and as the score decreases, it changes from light red to light blue, with dark blue representing the lowest weight.

4 Conclusion

In this work, we identify a fundamental limitation in current molecular property prediction methods: their reliance on singular, simplified representations of molecular knowledge, which restricts their generalizability across diverse tasks. To address this, we propose PyrMol, a novel heterogeneous pyramid framework that systematically integrates multidimensional prior knowledge. The core of PyrMol lies in its pyramid graph structure, which guides GNNs to capture multi-scale topological information. Furthermore, the introduced multi-source knowledge fusion module effectively balances the complementarity and redundancy of diverse expertise, emphasizing critical substructures. Augmented by a hierarchical contrastive learning strategy, our model significantly enhances the discriminative power of molecular representations.

Extensive experiments on ten molecular property prediction tasks demonstrate that PyrMol achieves superior performance over twelve state-of-the-art baselines, including advanced graph networks and subgraph augmentation methods. Remarkably, our end-to-end framework even rivals pretraining models by effectively integrating multi-perspective prior knowledge, establishing a reference paradigm for knowledge-driven approaches.

Despite its strong performance, PyrMol has limitations that point to future directions. The current framework incorporates three types of prior knowledge; however, molecular properties are influenced by a broader array of factors. Integrating additional reliable knowledge sources remains a promising avenue. Moreover, extending the pyramid to encompass three-dimensional molecular conformations could unlock further insights into structure-property relationships. We believe PyrMol offers a powerful and generalizable framework for molecular representation learning and anticipate it will contribute to accelerating AI-aided drug discovery.

References

1. Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y.A., Gomaa, M.M., Hassanien, A.E.: Deep learning in drug discovery: an integrative review and future challenges. Artificial Intelligence Review **56**(7), 5975–6037 (2023)
2. Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., et al.: Qsar modeling: where have you been? where are you going to? Journal of medicinal chemistry **57**(12), 4977–5010 (2014)
3. Degen, J., Wegscheid-Gerlach, C., Zaliani, A., Rarey, M.: On the art of compiling and using 'drug-like' chemical fragment spaces. ChemMedChem **3**(10), 1503–1507 (2008). `https://doi.org/https://doi.org/10.1002/cmdc.200800178`

4. Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., Wang, F.: A systematic study of key elements underlying molecular property prediction. Nature Communications **14**(1), 6395 (2023). https://doi.org/10.1038/s41467-023-41948-6
5. Gaudelet, T., Day, B., Jamasb, A.R., Soman, J., Regep, C., Liu, G., Hayter, J.B., Vickers, R., Roberts, C., Tang, J., et al.: Utilizing graph machine learning within drug discovery and development. Briefings in bioinformatics **22**(6) (2021)
6. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: International conference on machine learning. pp. 1263–1272. Pmlr (2017)
7. Hajiabolhassan, H., Taheri, Z., Hojatnia, A., Yeganeh, Y.T.: Funqg: Molecular representation learning via quotient graphs. Journal of chemical information and modeling **63**(11), 3275–3287 (2023)
8. Hamilton, W.L., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., 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. pp. 1024–1034 (2017)
9. Han, S., Fu, H., Wu, Y., Zhao, G., Song, Z., Huang, F., Zhang, Z., Liu, S., Zhang, W.: Himgnn: a novel hierarchical molecular graph representation learning framework for property prediction. Briefings in Bioinformatics **24**(5) (2023). https://doi.org/10.1093/bib/bbad305
10. Hinkson, I.V., Madej, B., Stahlberg, E.A.: Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. Frontiers in pharmacology **11**, 770 (2020)
11. Irwin, R., Dimitriadis, S., He, J., Bjerrum, E.J.: Chemformer: a pre-trained transformer for computational chemistry. Machine Learning: Science and Technology **3**(1), 015022 (2022)
12. Jaeger, S., Fulle, S., Turk, S.: Mol2vec: Unsupervised machine learning approach with chemical intuition. Journal of Chemical Information and Modeling **58**(1), 27–35 (2018). https://doi.org/10.1021/acs.jcim.7b00616
13. Jiang, Y., Jin, S., Jin, X., Xiao, X., Wu, W., Liu, X., Zhang, Q., Zeng, X., Yang, G., Niu, Z.: Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. Communications Chemistry **6**(1), 60 (2023). https://doi.org/10.1038/s42004-023-00857-x
14. Kengkanna, A., Ohue, M.: Enhancing property and activity prediction and interpretation using multiple molecular graph representations with mmgx. Communications Chemistry **7**(1), 74 (2024). https://doi.org/10.1038/s42004-024-01155-w
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017
16. Li, B., Lin, M., Chen, T., Wang, L.: Fg-bert: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. Briefings in Bioinformatics **24**(6) (2023)
17. Li, H., Zhang, R., Min, Y., Ma, D., Zhao, D., Zeng, J.: A knowledge-guided pre-training framework for improving molecular representation learning. Nature Communications **14**(1), 7568 (2023). https://doi.org/10.1038/s41467-023-43214-1
18. Song, Y., Zheng, S., Niu, Z., Fu, Z.H., Lu, Y., Yang, Y.: Communicative representation learning on attributed molecular graphs. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. pp. 2831–2838. ijcai.org. https://doi.org/10.24963/IJCAI.2020/392
19. Stiefl, N., Watson, I.A., Baumann, K., Zaliani, A.: Erg: 2d pharmacophore descriptions for scaffold hopping. Journal of Chemical Information and Modeling **46**(1), 208–220 (2006). https://doi.org/10.1021/ci050457y, https://doi.org/10.1021/ci050457y, pMID: 16426057
20. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M., Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R., Collins, J.J.: A deep learning approach to antibiotic discovery. Cell **180**(4), 688–702.e13 (2020). https://doi.org/10.1016/j.cell.2020.01.021, https://doi.org/10.1016/j.cell.2020.01.021, doi: 10.1016/j.cell.2020.01.021
21. Tang, X., Zhao, Q., Wang, J., Duan, G.: Molfcl: predicting molecular properties through chemistry-guided contrastive and prompt learning. Bioinformatics **41**(2) (2025). https://doi.org/10.1093/bioinformatics/btaf061
22. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: 6th International Conference on Learning Representations, ICLR 2018
23. Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V.: Moleculenet: a benchmark for molecular machine learning. Chemical Science **9**(2), 513–530 (2018). https://doi.org/10.1039/C7SC02664A
24. Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., Hsieh, C.Y., Hou, T.: Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. Nature Communications **14**(1), 2585 (2023). https://doi.org/10.1038/s41467-023-38192-3
25. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: 7th International Conference on Learning Representations, ICLR 2019

26. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R.: Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling **59**(8), 3370–3388 (2019). https://doi.org/10.1021/acs.jcim.9b00237
27. Zang, X., Zhao, X., Tang, B.: Hierarchical molecular graph self-supervised learning for property prediction. Communications Chemistry **6**(1), 34 (2023). https://doi.org/10.1038/s42004-023-00825-5
28. Zeng, X., Xiang, H., Yu, L., Wang, J., Li, K., Nussinov, R., Cheng, F.: Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. Nature Machine Intelligence **4**(11), 1004–1016 (2022)
29. Zhang, Y., Song, J., de la Fuente-Nunez, C.: Deep learning speeds the search for new antibiotic scaffolds. Nature Biotechnology (2025). https://doi.org/10.1038/s41587-025-02806-6