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

Yajie Li, Qichang Zhao, and Jianxin Wang, *Senior Member, IEEE,*

***Abstract*—Expert pharmaceutical chemists interpret molecular structures through a sophisticated cognitive hierarchy, transitioning from local functional moieties to spatial pharmacophores and, ultimately, to macroscopic pharmacological and physicochemical profiles. However, conventional Graph Neural Networks frequently overlook this high-level chemical intuition by treating molecules as single-scale atomic topology. To bridge this gap between human expertise and computational inference, we propose PyrMol, a knowledge-structured pyramid representation learning framework. By constructing heterogeneous hierarchical graphs, PyrMol orchestrates information flow across atomic, subgraph, and molecular levels. Crucially, the subgraph level systematically integrates three complementary expert views comprising functional groups, pharmacophores, and retrosynthetic fragments. To harmonize these explicit domain priors with implicit computational semantics, we introduce an adaptive Multi-source Knowledge Enhancement and Fusion module that dynamically balances their complementarity and redundancy. A Hierarchical Contrastive Learning strategy further ensures cross-scale semantic consistency. Empirical evaluations across ten benchmark datasets demonstrate that PyrMol outperforms 11 state-of-the-art baselines. Furthermore, its "plug-and-play" versatility provides a framework-agnostic performance boost for existing GNN architectures. PyrMol thus establishes a principled data-knowledge dual-driven paradigm for AI-aided Drug Discovery, effectively leveraging domain knowledge to catalyze advances in molecular property prediction.**

***Index Terms*—Molecular Property Prediction, Heterogeneous Message Passing Neural Network, Multi-source Knowledge Fusion, Contrastive learning.**

I. 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 [1]. 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 [3]. These models capitalize on various molecular featurization methods, including sequence-based [4], structure-based [5], and image-based descriptors [6].

Yajie Li, Qichang Zhao, and Jianxin Wang are with Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: jxwang@mail.csu.edu.cn).

For reproducibility, all code and data have been made available at https://github.com/lyj363636/PyrMol.

Notably, Graph Neural Networks (GNNs) stand out in learning molecular representations by propagating messages across chemical topologies [7]. Beyond foundational architectures (e.g., GCN [8], GIN [9], GAT [10]), specialized molecular networks like CMPNN [11] and AttentiveFP [12] have been explicitly tailored to model atom-bond interactions and extract higher-level substructural semantics, yielding superior predictive efficacy across standard evaluation datasets [13], [14], [15]. However, the performance of the aforementioned end-to-end models is often constrained by the limited availability of labeled molecular data. To address this bottleneck, pretraining frameworks (e.g., MolFCL [16], Himol [17], and KPGT [18]) have emerged as a pivotal school of thought to mitigate the reliance on extensive annotations, achieving remarkable progress by capturing rich structural semantics from unlabeled data.

Furthermore, expert chemists decode the causal relationships between molecular structures and properties through the identification of chemically significant substructures, rather than viewing molecules merely as low-level atom-bond networks [16]. Consequently, considerable research efforts [17], [18], [19], [20], [21] have pivoted towards incorporating chemical fragments to augment the predictive power of GNNs. Early substructure-enhanced methods typically rely on a single extraction paradigm. For instance, Hajiabolhassan et al. [20] proposed FunQG, a graph coarsening framework based specifically on functional groups to mitigate oversmoothing. Similarly, Jiang et al. [19] designed PharmHGT to extract vital substructures primarily derived from chemical reaction rules. While these methods have made strides in capturing localized semantics, their reliance on a single perspective inevitably creates an "information silo", inherently limiting the model's ability to capture the multi-view biochemical nature of molecules.

Fundamentally, distinct molecular properties are governed by different chemical dimensions. A comprehensive molecular profile requires integrating diverse expert knowledge: functional groups [21] for local physicochemical traits, pharmacophores [22] for spatial binding potentials, and retrosynthetic fragments [23] for macroscopic structural modularity. Crucially, their relative contributions vary significantly across downstream tasks; for instance, pharmacophores may dominate target affinity predictions, while functional groups contribute to analyse solubility. Therefore, predictions of different molecular properties are often reliable inferences derived from a combination of analyses from multiple perspectives.

To overcome this single-perspective limitation, recent approaches have explored incorporating diverse subgraphs. For example, Kengkanna et al. [24] investigated the applicabil-

ity of multiple subgraph views, such as junction trees and functional groups, within the MMGX framework. However, despite aggregating multiple perspectives, MMGX predominantly rely on naive aggregation techniques, failing to account for the intricate complementarity and inherent redundancy among heterogeneous domain knowledge, and struggling to dynamically align diverse chemical semantics. Consequently, this type of models often exhibit suboptimal generalizability across complex molecular property prediction tasks.

To break through this bottleneck and bridge the gap between sophisticated human expertise and computational inference, we propose PyrMol, a knowledge-structured pyramid graph framework for generalizable molecular property prediction. The main contributions of our work are as follows:

- We propose PyrMol, an end-to-end heterogeneous pyramid framework that orchestrates message passing across atomic, subgraph, and molecular levels. Its core topological design, Pyramid Molecular Graph, serves as a "plug-and-play" module, explicitly leveraging domain knowledge to comprehensively capture multi-scale molecular representations.
- We design an adaptive multi-source knowledge enhancement and fusion module to seamlessly integrate disparate streams of chemical expertise. By dynamically balancing knowledge complementarity and redundancy, this module effectively mitigates the poor generalization and robustness issues stemming from single-perspective representations.
- We develop a hierarchical contrastive learning strategy across multiple representation scales (atoms, subgraphs, and molecules). This cross-scale alignment enforces semantic consistency and maximizes the discriminative power of the learned embeddings.

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.

II. Method

As depicted in Fig.1, the framework of PyrMol contains three core components. First, we construct a Heterogeneous Pyramid Molecular Graph (Fig.1.A) to hierarchically abstract a molecule into Atom-, Sub-, and Mol-level topologies. Second, a knowledge-guided heterogeneous Message Passing Neural Network (MPNN) with an adaptive fusion module (Fig.1.B) is deployed to update multi-scale representations. This module dynamically integrates multi-source knowledge into a unified Sub-Fuse embedding, aligning its dimension with Atom- and Mol-embeddings for cross-scale interactions. Third, a hierarchical contrastive learning strategy (Fig.1.C) mutually regularizes and aligns semantic representations across the three levels. Ultimately, the aligned features are concatenated for downstream property prediction via a fully connected network.

A. Heterogeneous Pyramid Message Passing Neural Network

1) Pyramid Molecular Graph Construction: To explicitly guide GNNs in capturing hierarchical substructures, we construct a heterogeneous Molecular Pyramid Graph comprising three topological levels (Fig.1.a). The base level consists of atom nodes V_A and bond edges E_A, representing the fundamental atomic topology. The middle level encompasses subgraph nodes V_{S_i} (representing key structural or functional motifs) and their relational edges. Specifically, E_{AS_i} connects atoms to their corresponding subgraphs to direct explicit bottom-up feature aggregation, while E_{S_i} connects adjacent or overlapping subgraphs within the i-th knowledge type. The top level contains a single global molecular node V_M connected to all subgraphs via E_{S_iM}, designed to aggregate localized semantics into a comprehensive global representation.

To populate the middle level, we integrate three complementary types of domain knowledge—functional groups (F), pharmacophores (P), and retrosynthetic fragments (R)—to parse molecules into distinct meaningful fragments. Specifically, we utilize RDKit to extract chemical functional groups (V_F); employ the ErG algorithm [22] to identify biophysical pharmacophoric subgraphs (V_P, encompassing six non-covalent interaction potentials such as H-bond donors/acceptors); and apply predefined BRICS rules [23] to generate retrosynthetically guided reaction fragments (V_R). Consequently, the complete pyramid graph is formalized as $G = \{V, E\}$, where $V = \{V_A, V_P, V_F, V_R, V_M\}$ and $E = \{E_A, E_P, E_F, E_R, E_{AP}, E_{AF}, E_{AR}, E_{PM}, E_{FM}, E_{RM}\}$. Each node features are combined chemical features(detailed in Supplementary Table I) and computational features(initialed by Mol2Vec[25]). Initial node features are rapidly computed *in silico* via RDKit.

TABLE I
NODE FEATURES 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, the explicit bonds connected to the atom, hybridization type.
Pharmacophore	6	H-bond donor, H-bond acceptor, positive, negative, hydrophobic, aromatic[22].
Functional Group	115	Predefined types of edges, rings, functional groups.
Retrosynthetic Fragments	27	BRICS reaction type features[23].
Molecule	300	The average of atoms.

2) Heterogeneous Message Passing Neural Networks.: To endow the model with a robust baseline understanding of chemical semantics prior to hierarchical message passing, we initialize the intrinsic node attributes utilizing Mol2Vec embeddings [25]. It is imperative to note that within the PyrMol framework, these external embeddings serve strictly as static initial node features F_i rather than requiring a computationally expensive pretraining phase. 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.

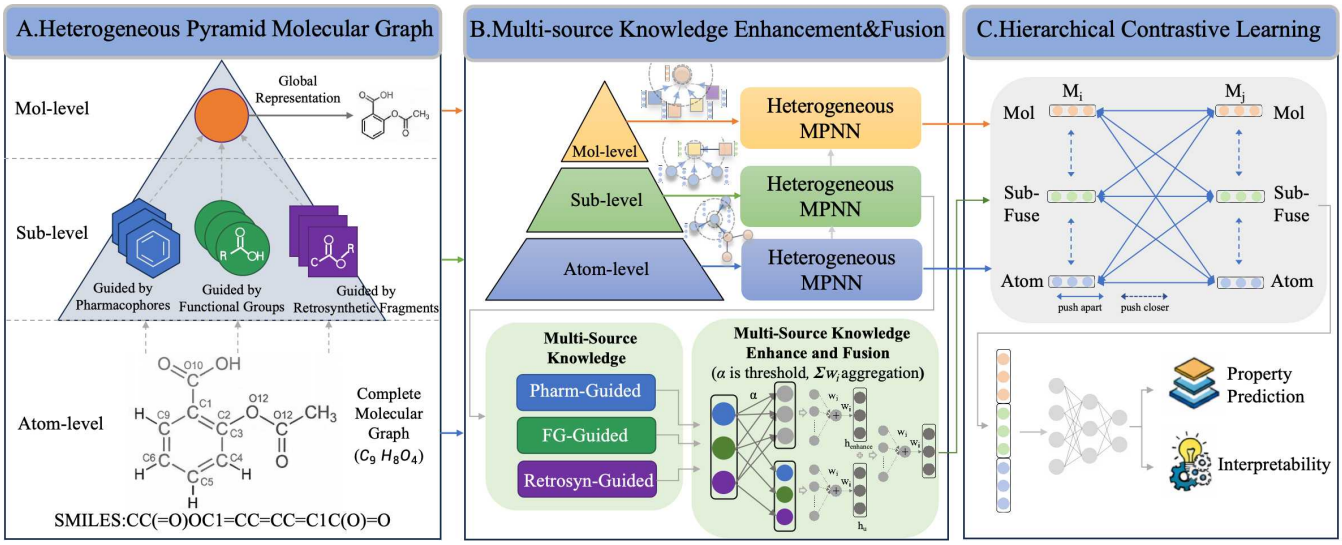


Fig. 1. **Illustration of Knowledge-Structured Pyramid Graph Framework(PyrMol).** PyrMol contains three main modules: A. Heterogeneous Pyramid Graph; B. Multi-Source Knowledge Enhance and Fusion; C. Hierarchical Contrast.

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

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.

The architectural selection of our heterogeneous MPNN is explicitly driven by the varying nature of chemical message passing across scales. At the atomic and substructural levels, the heterogeneous graph formed by atom nodes V_A and subgraph nodes (V_P, V_R, V_F) undergoes feature updates via GraphSAGE [26]. This leverages its inductive learning capability to efficiently aggregate localized topological environments and intra-subgraph connectivity. In contrast, at the molecular level, a Graph Attention Network (GAT) [10] is utilized to aggregate signals from these diverse substructure nodes to the global molecular node V_M. This design choice is motivated by the chemical intuition that different substructures (e.g., a critical pharmacophore versus a generic carbon backbone) contribute unequally to the global molecular property, allowing the attention mechanism to dynamically learn these differential contributions.Formally, 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:

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

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

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

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 retrosynthetic fragment feature $h_R \in \mathbb{R}^k$ are obtained.

B. Multi-source Knowledge Enhancement and Fusion Module

As shown in Fig.1.b., 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 $h_F \in \mathbb{R}^k$, $h_P \in \mathbb{R}^k$, and $h_R \in \mathbb{R}^k$, we calculate the similarity between pairs of them.

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

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 s_{ij} is greater than threshold $\alpha = 0.3$. If it is, we generate their mutual key feature vector $k_{ij} \in \mathbb{R}^k$ that are essential for them.

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

where the key feature matrix $K \in \mathbb{R}^{3 \times d}$ is composed by $\{k_{FP}, k_{FR}, k_{PR}\}$.

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

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.

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

By mutually regularizing three streams of explicit domain priors (V_F, V_P, V_R) with the implicit continuous computational semantics, this module establishes a powerful data-knowledge dual-driven paradigm. This principled synergy between discrete symbolic rules and data-driven representations dynamically highlights consensus features and effectively dismantles information silos, yielding a robust and comprehensive molecular embedding.

C. Hierarchical Contrastive Learning Strategy and Prediction

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. In experiments, we set the hyperparameter λ to 0.2. $loss_{task}$ is Binary Cross-Entropy Loss for classification tasks and Mean Square Error loss for regression tasks.

III. EXPERIMENTS AND DISCUSSION

A. Datasets and Baselines

To evaluate the effectiveness and generalizability of our proposed model, we conducted extensive experiments on ten benchmark datasets from MoleculeNet [27]. These include seven classification tasks—Blood-Brain Barrier Permeability (BBBP), BACE, HIV, ClinTox, Tox21, SIDER, and ToxCast—and three regression tasks—ESOL, FreeSolv, and Lipophilicity. Model performance is primarily assessed using the Area Under the Receiver Operating Characteristic curve (AUC) for classification and Root Mean Square Error (RMSE) for regression. To provide a more granular performance analysis, we also report a comprehensive suite of auxiliary metrics, including Accuracy, Precision, Recall, F1-Score, and AUPR for classification, alongside MSE, MAE, and R^2 for regression. Detailed dataset descriptions are provided in Supplementary Section A.

To demonstrate the competitive advantages of PyrMol, we benchmarked it against 11 state-of-the-art (SOTA) methods. These baselines encompass a diverse range of architectures: traditional graph neural networks (GCN [8], GIN [9], and GAT [10]), specialized molecular graph networks (AttentiveFP [12] and CMPNN [11]), subgraph-enhanced models (HiMol_w/oPretrain [18], PharmHGT [19], and MMGX [21]), and pretraining-based frameworks (HiMol [18], KPGT [28], and MolFCL [17]). Notably, HiMol_w/oPretrain [18] is included as a non-pretrained variant to isolate the architectural contributions of the HiMol framework.

B. Implementation details

To ensure a rigorous and unbiased evaluation across benchmark datasets, we employ a random scaffold splitting strategy [29]. Unlike the conventional scaffold splitting approach proposed in MoleculeNet [27], which typically sorts scaffolds by frequency and yields deterministic partitions, this strategy introduces stochasticity into the data segmentation. As noted by Deng et al. [29], deterministic splitting procedures can inadvertently favor specific model architectures, thereby inflating perceived generalizability. To mitigate this risk, we eliminate frequency-based sorting and instead shuffle the scaffolds using a random seed. This strategy maintains the structural dissimilarity between training and test sets—crucial for assessing out-of-distribution performance—while providing a more robust partitioning scheme.

All datasets are partitioned into training, validation, and test sets with a ratio of 8:1:1. To achieve statistical significance, we perform 25 independent trials using distinct random seeds. For a fair comparison, all baseline models are implemented via their official repositories using the optimal hyperparameters reported in their respective original studies. Importantly, to ensure an absolutely rigorous and fair comparison and eliminate any bias stemming from data partitioning, all state-of-the-art baseline models were meticulously retrained and evaluated under the exact same random scaffold splitting criterion and random seeds. Our framework is implemented in PyTorch and executed on an NVIDIA Tesla A100 (40GB) GPU, with molecular descriptors (node and edge features) generated using the RDKit library.

C. Comparison with Baselines

To comprehensively evaluate our framework, we benchmarked PyrMol against 11 state-of-the-art baselines across 10 distinct molecular property datasets. To summarize the global performance, the average ranking (Avg Rank) across all tasks is employed, as presented in Table II. Overall, PyrMol demonstrates commanding superiority, achieving the best Average Rank (1.8) among all 13 models. Specifically, it achieves the top performance on 6 out of 10 tasks, while maintaining highly competitive results on the remaining benchmarks.

1) Superiority over Conventional GNNs: When compared to traditional end-to-end graph networks (i.e., GCNs, GINs, GATs) and specialized molecular architectures (AttentiveFP, CMPNN), PyrMol exhibits substantial performance improvements. For instance, on the classification tasks, PyrMol achieves remarkable AUC scores of 0.949 on BBBP, 0.871 on BACE, and 0.811 on HIV. These represent absolute

TABLE II
PERFORMANCE OF PYRMOL AND BASELINE METHODS ON MOLECULAR PROPERTY PREDICTION.

Models	Classification(AUC↑)			Multi-task Classification(AUC↑)				Regression(RMSE↓)			Avg
	BBBP	BACE	HIV	ClinTox	SIDER	Tox21	ToxCast	FreeSolv	Lipophilicity	ESOL	Rank↑
GCNs [8]	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)	7.7
GINs [9]	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)	7.7
GATs [10]	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)	8
AttentiveFP [12]	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)	9.4
CMPNN [11]	0.851 (0.128)	0.840 (0.052)	0.746 (0.038)	0.842 (0.045)	0.535 (0.025)	0.795 (0.021)	0.654 (0.016)	1.784 (0.427)	0.627 (0.027)	0.777 (0.085)	6.9
Himol_w/oPretrain [18]	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)	8.8
PharmHGT [19]	0.932 (0.029)	0.834 (0.044)	0.764 (0.040)	0.866 (0.057)	0.619 (0.029)	**0.834** (0.011)	0.717 (0.010)	2.361 (0.621)	0.707 (0.037)	0.956 (0.200)	4.1
MMGX [21]	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)	9.5
HiMol [18]	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)	6.9
KPGT [28]	0.922 (0.025)	0.855 (0.033)	0.791 (0.036)	0.854 (0.064)	**0.635** (0.015)	**0.834** (0.015)	0.704 (0.009)	2.130 (0.660)	**0.596** (0.018)	0.943 (0.143)	3.2
MolFCL [17]	0.872 (0.070)	0.833 (0.046)	0.796 (0.037)	0.862 (0.077)	0.598 (0.033)	0.808 (0.020)	0.711 (0.016)	1.777 (0.494)	0.638 (0.029)	**0.769** (0.091)	4.2
PyrMol	**0.949** (0.020)	**0.871** (0.036)	**0.811** (0.041)	**0.878** (0.034)	0.622 (0.026)	**0.834** (0.020)	**0.722** (0.009)	**1.697** (0.400)	0.668 (0.040)	0.865 (0.120)	**1.6**

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.

improvements of 6.8%, 3.1%, and 4.5%, respectively, over the strongest baselines in these categories. Furthermore, on the FreeSolv regression task, PyrMol reduces the RMSE to 1.697. These significant improvements empirically validate that explicitly orchestrating molecular topologies into a hierarchical pyramid graph—rather than relying solely on single-scale atomistic topology—profoundly enhances the model's capacity to capture complex chemical environments.

2) Advantage over Subgraph-Enhanced Models: In contrast to subgraph-enhanced models, PyrMol consistently surpasses single-perspective models like PharmHGT. More importantly, PyrMol decisively outperforms MMGX, a competitive model that also utilizes multiple subgraphs but lacks an adaptive fusion strategy (Avg Rank: 1.8 vs. 10.1). This comparison highlights the critical advantage of our Multi-source Knowledge Enhancement and Fusion Module. Merely extracting diverse substructures is insufficient, and dynamically balancing their complementarity and redundancy through principled heterogeneous message passing is essential for optimal feature extraction.

3) Efficiency against Pretraining Models: Furthermore, we compared PyrMol against state-of-the-art pretraining frameworks (e.g., KPGT, MolFCL, HiMol). Remarkably, PyrMol achieves a superior overall ranking without requiring computationally exhaustive, massive-scale pretraining phases. By strategically synergizing explicit domain priors with simple implicit computational semantics (Mol2Vec), our data-knowledge dual-driven paradigm emerges as a highly efficient alternative to complex pretraining strategies.

Additional evaluation metrics, e.g., Accuracy for classification and MSE for regression, are provided in Supplementary Section B.

D. Validation of the Pyramid Molecular Graph

To validate the "plug-and-play" versatility of the proposed Pyramid Molecular Graph (PMG) and further demonstrate its intrinsic topological value, we explicitly replaced the native atomic graphs (or pre-defined hierarchical graphs) of five established baselines with our PMG module. The evaluated baselines encompass both classical foundational GNNs (GCNs, GINs, GATs) and advanced subgraph-augmented architectures (HiMol_w/oPretrain, PharmHGT).

We evaluated these PMG-enhanced variants across six representative datasets. As comprehensively visualized in the relative performance gain heatmap (Fig.2 b) and performance comparison bar graph(Fig.2 a). The specific details showed in Supplementary Table S11, the integration of the PMG module yielded performance improvements in 24 out of the 30 experimental evaluations.

Notably, foundational architectures that inherently lack explicit domain knowledge experienced the most profound benefits. GCNs, the most fundamental architecture evaluated, achieved consistent gains across all six datasets with an impressive average performance improvement of 7.06%. Similarly, GINs and GATs secured average overall boosts of 5.54% and 4.73%, respectively. The impact of PMG acting as a transformative structural prior is most acutely observed on the FreeSolv dataset, where PMG integration catapulted the performance of GINs, GCNs, and GATs by exceptional margins of 31.42% (i.e., $(3.046-2.089)/3.046$), 23.85%, and 23.42%, respectively. This underscores the capacity of the pyramid topology to empower simple architectures to extract highly expressive, expert-level structural features.

Furthermore, leading subgraph-enhanced baselines also exhibited robust performance enhancements, with the PMG-

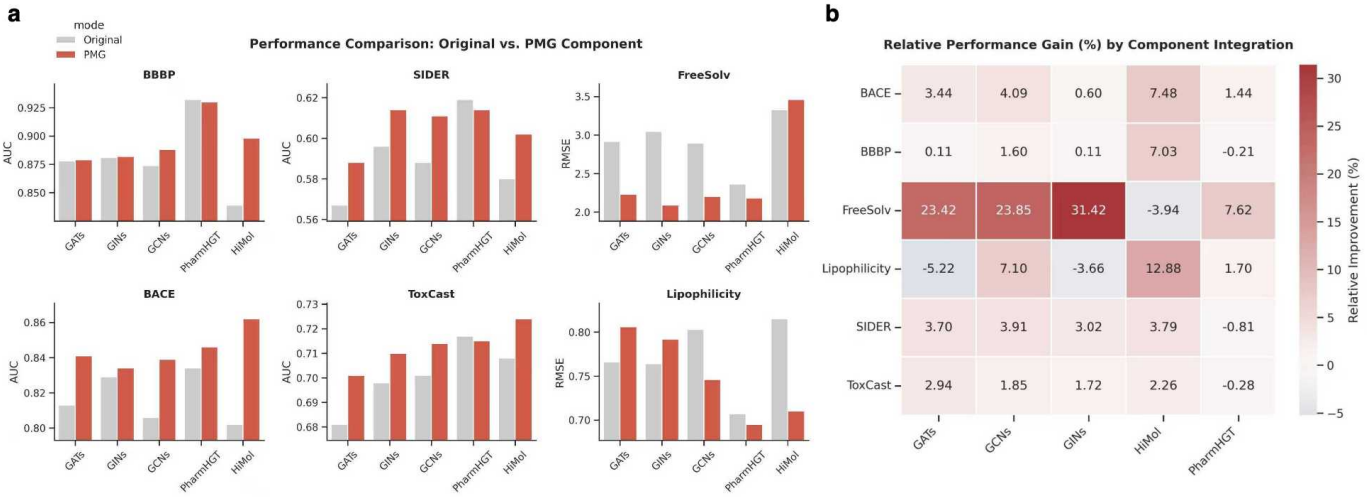


Fig. 2. **Relative Performance Gain of the baselines with the Pyramid Molecular Graph.** a is performance comparison bar graph between original models and the models added PMG Component. b is the relative performance gain heatmap

augmented versions of HiMol_w/oPretrain and PharmHGT achieving average gains of 4.92% and 1.58% over their original implementations.

A particularly compelling observation emerges when evaluating the necessity of massive pretraining. As detailed in Supplementary Table S12, substituting the native topology with our PMG module—without utilizing any pretraining phase (HiMol_PMG)—actually surpasses the fully pretrained HiMol model on five out of six benchmarks (BBBP, BACE, SIDER, ToxCast, and Lipophilicity). For instance, on the BBBP dataset, HiMol_PMG achieves an AUC of 0.898, outperforming both Himol_w/oPretrain (0.839) and the pretrained Himol (0.889). This finding implies that explicitly guiding models with structured, multi-scale domain priors is a highly effective strategy, offering a computationally lightweight complement or alternative to having models implicitly learn these semantics through extensive pretraining.

Collectively, these robust statistical averages and the profound performance shifts against pretrained models validate that our proposed PMG acts as a universal, framework-agnostic enhancement mechanism. By explicitly orchestrating multi-level node message passing, PMG fundamentally elevates the feature learning efficiency of diverse graph neural networks.

E. Effectiveness Analysis of Multi-source Knowledge Enhancement and Fusion

To rigorously validate the strict necessity of integrating diverse chemical expertise, we conducted an ablation study exploring various combinations of prior knowledge. Since the atomic graph (A) provides the fundamental molecular topology, we established it as the universal base graph. Subsequently, we evaluated seven distinct subgraph-enhanced modes by systematically augmenting the atomic graph with functional groups (F), pharmacophores (P), and retrosynthetic fragments (R). The performance across diverse classification and regression benchmarks is summarized in Table III.

Evaluating the single-knowledge enhancements (A+R, A+F, and A+P) reveals a clear task-dependent preference. For instance, on the BBBP dataset, incorporating functional groups (A+F) yields a significantly better AUC (0.928) compared to pharmacophores (A+P, 0.908) or reaction rules (A+R, 0.909). Conversely, A+P and A+R demonstrate unique advantages on multi-classification tasks. This observation perfectly aligns with our core hypothesis: distinct molecular properties are governed by varying physical and chemical mechanisms. Consequently, relying on any single-perspective subgraph is inherently insufficient for universal generalization across diverse biochemical tasks.

Furthermore, as progressively more knowledge sources are integrated, the predictive performance consistently improves, eventually reaching the optimal performance of the fully integrated A+R+F+P model. As shown in Table III, the full combination achieves the highest predictive capability across almost all metrics, including the lowest RMSE on FreeSolv (2.475) and Lipophilicity (0.656), as well as the highest AUC on BBBP (0.934) and SIDER (0.559).

TABLE III
MODEL PERFORMANCE OF MULTIPLE GRAPH COMBINATION MODELS.

Subgraphs	Classification(AUC↑)				Regression(RMSE↓)	
	BBBP	BACE	SIDER	ToxCast	FreeSolv	Lipophilicity
A	0.878 (0.037)	0.813 (0.048)	0.566 (0.025)	0.671 (0.014)	2.916 (0.739)	0.766 (0.051)
A+R	0.909 (0.034)	0.852 (0.053)	0.553 (0.024)	0.695 (0.014)	2.874 (0.770)	0.703 (0.025)
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, retrosynthetic fragments, functional groups, and pharmacophores, respectively.

Models	PMG	M2V	EF	CL
PyrMol	√	√	√	√
Variant 1	√	√	√	×
Variant 2	√	√	×	×
Variant 3	√	×	√	√
Variant 4	√	×	×	×
Variant 5	×	×	×	×

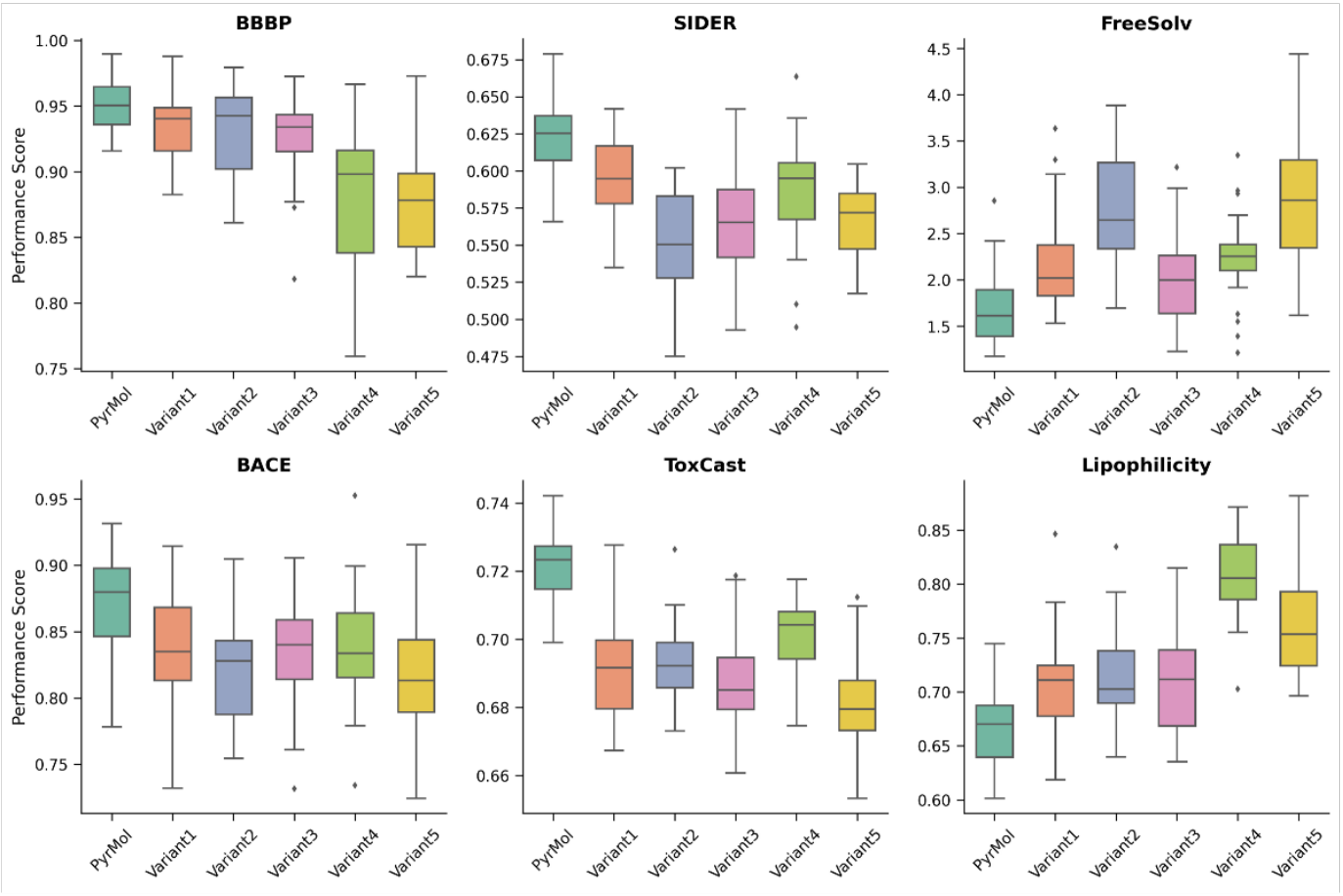


Fig. 3. **The results of ablation study on six datasets.** The figure contains models definition details introduction and performance boxplot figure.

These results underscore that the integration of functional groups (F), pharmacophores (P), and retrosynthetic fragments (R) priors is not merely additive, but highly synergistic. Equipping the model with this comprehensive, complementary knowledge base is essential

F. Ablation Study

To rigorously evaluate the contribution of each component in PyrMol, we conducted a comprehensive ablation study across six benchmark datasets. We designed five variants to analyze the specific effects of the Pyramid Molecular Graph (PMG), Hierarchical Contrastive Learning Strategy (CL), Multi-source Knowledge Enhancement and Fusion Module (EF), and the integration of Mol2Vec features (M2V). The results are shown in Fig.3.

Comparing Variant 4 with the Variant 5 demonstrates the intrinsic superiority of our topological design. Even without auxiliary features or learning strategies, the Pyramid Hierarchical Graph construction significantly outperforms the Variant 5, particularly in classification tasks. Besides, on the FreeSolv dataset, Variant 4 reduces the RMSE from 2.916 to 2.233. This indicates that our Pyramid Hierarchical Graph construction effectively captures multi-scale structural information inherent in molecular graphs, providing a robust geometric foundation.

Notably, the complete PyrMol framework substantially outperforms both Variant 2 and Variant 3. The significant performance drop observed in both partial configurations indicates a powerful synergistic effect. It demonstrates that simply supplementing initial node features via Mol2Vec is insufficient; rather, it provides a foundational semantic space that strictly requires our EF and CL modules to be effectively aligned,

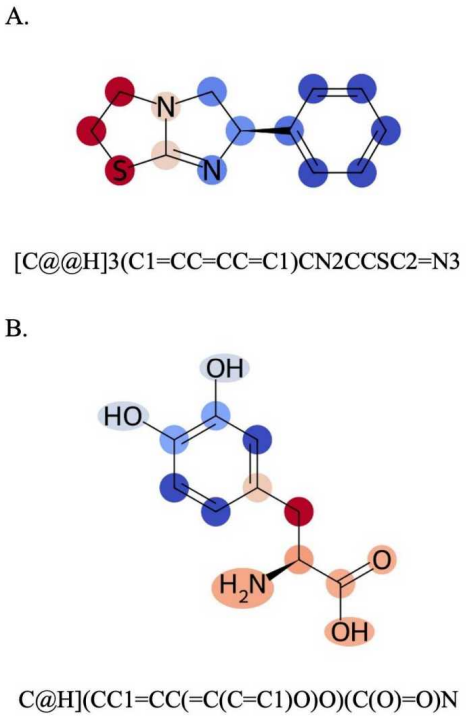


Fig. 4. **Interpretability analysis of attention scores on the BBBP dataset.** Dark red indicates the atom have the highest attention score, light red the next highest, and as the score decreases, it changes from light red to light blue, with dark blue representing the lowest weight.

refined, and translated into superior predictive performance.

The comparison between Variant 1 and PyrMol highlights that the CL module consistently refines the feature space across all six tasks, yielding universal performance improvements (e.g., pushing the AUC on BBBP from 0.936 to 0.949). Finally, the full PyrMol model achieves the best performance across all metrics. While Mol2Vec provides a useful initialization, the dominant performance gains are driven by the synergy of our Pyramid architecture and learning strategies, confirming that our proposed modules are the primary drivers of PyrMol's performance.

G. Interpretation Results

A key advantage of PyrMol is its inherent interpretability. Because PyrMol utilizes the GAT architecture to aggregate features from substructure nodes to the global molecular node, and the hierarchical mapping between atoms and substructures is deterministically defined within the pyramid graph, the model can explicitly trace and project learned attention scores back to individual atoms. To validate this capability, we visualize the atomic attention weights for representative molecules from the BBBP dataset. According to the visualization scheme, dark red indicates the highest attention score, transitioning through light red to light blue, with dark blue representing the lowest weight.

As illustrated in Fig. 4A, the first molecule ($[C@@H]3(C1 = CC = CC = C1)CN2CCSC2 = N3$) features a rigid planar structure formed by two fused heterocycles attached to a benzene ring. Chemically, the sulfur atom in the thiazole ring exhibits lower polarity and stronger hydrophobicity than typical oxygen analogues, thereby positively contributing to BBB permeability. Crucially, PyrMol captures this phenomenon, assigning the highest attention weight (visualized as dark red) to the sulfur atom. Conversely, the sp^2 hybridized nitrogen atom acts as a strong hydrogen bond acceptor, typically impeding BBB penetration. The model accurately identifies this resistance point, assigning it a negatively correlated, lower attention score (visualized as blue).

As shown in Fig. 4B, the second molecule ($C[@H](CC1 = CC(= C(C = C1)O)O)(C(O) = O)N$) comprises an amino acid backbone, an aromatic side chain, catechol hydroxyl groups, and a methylene bridge. PyrMol correctly recognizes the amino acid backbone ($H_2N - CH - COOH$) with a high correlation for BBB crossing (visualized as light red, indicating the next highest attention). Meanwhile, the crucial hydrophobic methylene bridge ($-CH_2-$) connecting the main and side chains is rightly highlighted with the highest attention score (dark red).

Collectively, these visualizations confirm that the attention mechanisms within PyrMol align with sophisticated pharmaceutical intuition. By focusing on atoms and substructures that dictate macroscopic properties, PyrMol provides highly reliable and explainable clues for downstream molecular optimization.

IV. Conclusion

In this work, we propose PyrMol, a knowledge-structured heterogeneous pyramid framework that pioneers a data-knowledge dual-driven paradigm for molecular representation learning. Moving beyond conventional single-scale atomic topology, PyrMol systematically mirrors the cognitive hierarchy of expert chemists. Its core pyramid graph orchestrates message passing across multi-scale chemical environments, while the adaptive Multi-source Knowledge Enhancement and Fusion module dynamically balances the complementarity and redundancy of heterogeneous domain priors. Complemented by a hierarchical contrastive learning strategy, PyrMol significantly amplifies the discriminative power of molecular embeddings. Extensive evaluations across ten benchmarks demonstrate that PyrMol consistently outperforms twelve state-of-the-art baselines. Remarkably, without relying on computationally exhaustive pretraining phases, our end-to-end framework establishes a highly efficient and robust reference paradigm, proving the immense potential of explicitly integrating multi-perspective structural knowledge.

While PyrMol demonstrates competitive performance on 2D molecular graphs, there are several directions for future improvement. Currently, the framework integrates three specific types of domain knowledge; however, incorporating a broader range of reliable features could further improve molecular representations. Furthermore, extending the current pyramid structure to encompass three-dimensional (3D) molecular conformations is a necessary next step to better capture stereochemical structure-property relationships. We hope that PyrMol can serve as a practical and effective framework for molecular representation learning, providing useful computational support for AI-aided drug discovery.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. U25A20449, 62272309 and U22A2041). This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

REFERENCES

[1] I. V. Hinkson, B. Madej, and E. A. Stahlberg, "Accelerating therapeutics for opportunities in medicine: A paradigm shift in drug discovery," *Frontiers in pharmacology*, vol. 11, p. 770, 2020.

[2] A. Cherkasov et al., "Qsar modeling: Where have you been? where are you going to?" *Journal of medicinal chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[3] H. Askr, E. Elgeldawi, H. Aboul Ella, Y. A. Elshaier, M. M. Gomaa, and A. E. Hassanien, "Deep learning in drug discovery: An integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 5975–6037, 2023.

[4] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: A pre-trained transformer for computational chemistry," *Machine Learning: Science and Technology*, vol. 3, no. 1, p. 015 022, 2022.

[5] B. Li, M. Lin, T. Chen, and L. Wang, "Fg-bert: A generalized and self-supervised functional group-based molecular representation learning framework for properties prediction," *Briefings in Bioinformatics*, vol. 24, no. 6, 2023.

[6] X. Zeng et al., "Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 1004–1016, 2022.

[7] T. Gaudelet et al., "Utilizing graph machine learning within drug discovery and development," *Briefings in bioinformatics*, vol. 22, no. 6, 2021.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations*, 2017.

[9] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" In *7th International Conference on Learning Representations*, 2019.

[10] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations*, 2018.

[11] Y. Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, and Y. Yang, "Communicative representation learning on attributed molecular graphs," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2000, pp. 2831–2838.

[12] K. Yang et al., "Analyzing learned molecular representations for property prediction," *Journal of Chemical Information and Modeling*, vol. 59, no. 8, pp. 3370–3388, 2019, ISSN: 1549-9596.

[13] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, Pmlr, 2017, pp. 1263–1272.

[14] J. M. Stokes et al., "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, 688–702.e13, 2020, doi: 10.1016/j.cell.2020.01.021, ISSN: 0092-8674.

[15] Y. Zhang, J. Song, and C. de la Fuente-Nunez, "Deep learning speeds the search for new antibiotic scaffolds," *Nature Biotechnology*, 2025, ISSN: 1546-1696.

[16] Z. Wu et al., "Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking," *Nature Communications*, vol. 14, no. 1, p. 2585, 2023, ISSN: 2041-1723.

[17] X. Tang, Q. Zhao, J. Wang, and G. Duan, "Molfcl: Predicting molecular properties through chemistry-guided contrastive and prompt learning," *Bioinformatics*, vol. 41, no. 2, 2025, ISSN: 1367-4811.

[18] X. Zang, X. Zhao, and B. Tang, "Hierarchical molecular graph self-supervised learning for property prediction," *Communications Chemistry*, vol. 6, no. 1, p. 34, 2023, ISSN: 2399-3669.

[19] Y. Jiang et al., "Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction," *Communications Chemistry*, vol. 6, no. 1, p. 60, 2023, ISSN: 2399-3669.

[20] H. Hajiabolhassan, Z. Taheri, A. Hojatnia, and Y. T. Yeganeh, "Funqg: Molecular representation learning via quotient graphs," *Journal of chemical information and modeling*, vol. 63, no. 11, pp. 3275–3287, 2023.

[21] A. Kengkanna and M. Ohue, "Enhancing property and activity prediction and interpretation using multiple molecular graph representations with mmgx," *Communications Chemistry*, vol. 7, no. 1, p. 74, 2024, ISSN: 2399-3669.

[22] N. Stiefl, I. A. Watson, K. Baumann, and A. Zaliani, "Erg: 2d pharmacophore descriptions for scaffold hopping," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 208–220, 2006.

[23] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey, "On the art of compiling and using 'drug-like' chemical fragment spaces," *ChemMedChem*, vol. 3, no. 10, pp. 1503–1507, 2008.

[24] Z. Zhang, J. Guan, and S. Zhou, "Fragat: A fragment-oriented multi-scale graph attention model for molecular property prediction," *Bioinformatics*, vol. 37, no. 18, pp. 2981–2987, Mar. 2021, ISSN: 1367-4803.

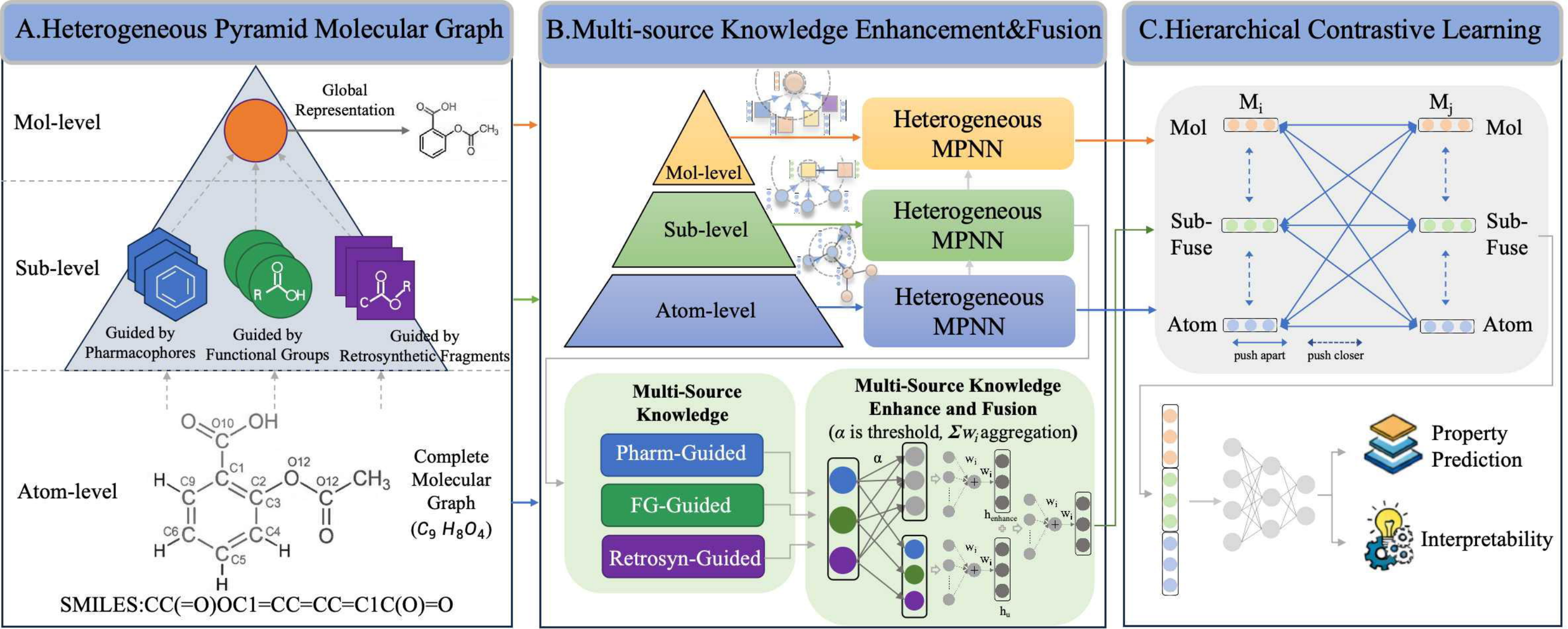
[25] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised machine learning approach with chemical intuition," *Journal of Chemical Information and Modeling*, vol. 58, no. 1, pp. 27–35, 2018.

[26] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1024–1034.

[27] Z. Wu et al., "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018, ISSN: 2041-6520.

[28] H. Li, R. Zhang, Y. Min, D. Ma, D. Zhao, and J. Zeng, "A knowledge-guided pre-training framework for improving molecular representation learning," *Nature Communications*, vol. 14, no. 1, p. 7568, 2023, ISSN: 2041-1723.

[29] J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras, and F. Wang, "A systematic study of key elements underlying molecular property prediction," *Nature Communications*, vol. 14, no. 1, p. 6395, 2023, ISSN: 2041-1723.


A.Heterogeneous Pyramid Molecular Graph
Mol-level
Global Representation
Sub-level
Guided by Pharmacophores
Guided by Functional Groups
Guided by Retrosynthetic Fragments
Atom-level
Complete Molecular Graph
($C_9H_8O_4$)
SMILES:CC(=O)OC1=CC=CC=C1C(O)=O
B.Multi-source Knowledge Enhancement&Fusion
Mol-level
Sub-level
Atom-level
Heterogeneous MPNN
Heterogeneous MPNN
Heterogeneous MPNN
Multi-Source Knowledge
Pharm-Guided
FG-Guided
Retrosyn-Guided
Multi-Source Knowledge Enhance and Fusion
(α is threshold, Σw_i aggregation)
C.Hierarchical Contrastive Learning
M_i
M_j
Mol
Sub-Fuse
Atom
push apart
push closer
Property Prediction
Interpretability