HIERARCHICAL MIXTURE OF TOPOLOGICAL EXPERTS FOR MOLECULAR PROPERTY PREDICTION

Kiwoong Yoo
AIGEN Sciences
kiwoong.yoo@aigensciences.com

Jaewoo Kang*
AIGEN Sciences, Korea University
kangj@korea.ac.kr

ABSTRACT

Molecular property prediction enables rapid identification of promising drug candidates by forecasting key attributes such as bioactivity and toxicity. The relationship between molecular structure and properties spans multiple scales—from individual atoms to functional groups to the overall molecular framework. Depending on the property task and the target molecule's scaffold, prediction may require focusing on specific substructures or the entire molecular configuration. This observation suggests that selectively attending to relevant structural features at different scales can improve prediction accuracy. In this light, we propose HierMolMoE, a hierarchical mixture-of-experts framework that learns specialized predictive models at three natural granularities of molecular graphs: atom-level, motif-level, and global-level. Our model integrates expert networks at each level with a high-level gating mechanism, and each expert is tailored to capture the unique topological semantics of molecular groups sharing similar scaffolds. Experiments on benchmark datasets demonstrate that HierMolMoE outperforms existing GNN-based mixture-of-experts approaches for molecular property prediction, highlighting its ability to learn robust structure–property relationships across scales.

1 INTRODUCTION

Molecular property prediction is essential for drug discovery, enabling the rapid identification of candidates by forecasting properties such as bioactivity, toxicity, and pharmacokinetics (Shen & Nicolaou (2019)). However, property prediction is challenging because it depends on a multitude of factors that vary by task, with molecular features relevant at different scales. For example, blood-brain barrier permeability might hinge on local features like hydrogen bonding capacity in some molecules, while in others, global attributes such as molecular weight and polar surface area are decisive (Kadry et al. (2020)). Similarly, enzyme inhibition depends on both key pharmacophoric elements and broader structural context (Roy & Roy (2009)). These examples underscore the need for models that can adaptively capture both local and global molecular characteristics.

Graph Neural Networks (GNNs) have emerged as powerful tools for encoding molecular graphs into multi-scale representations (Yang et al. (2019)). Many methods have enhanced GNNs by exploiting recurring subgraphs (motifs, Zhang et al. (2021); Peng et al. (2020); Wu et al. (2023)) or by emphasizing specific topological configurations during aggregation (Chen & Gel (2023); Baek et al. (2021); Islam et al. (2023); Ying et al. (2018)). However, most approaches rely on a single GNN operating uniformly over the graph, which can obscure scale-specific details (Rusch et al. (2023)).

Mixture-of-Experts (MoE) architectures offer an attractive alternative by employing multiple specialized predictors (Kim et al. (2023); Yao et al. (2023); Soares et al. (2024)). Yet, existing MoE approaches for molecular prediction treat the molecule as a whole, neglecting its inherent hierarchical structure.

To address these challenges, we propose HierMolMoE, a hierarchical mixture-of-experts framework that learns specialized predictors at three natural scales of molecular graphs: atom-level interactions, substructure-level motifs, and global molecular representations. Each level further subdivides experts based on distinct scaffolds, enabling the model to capture both fine-grained and overall topo-

*Corresponding author.

logical features. By dynamically assigning molecules to one or more experts according to both local and scaffold-specific cues, HierMolMoE effectively models the multi-scale, scaffold-dependent nature of structure–property relationships.

Our main contributions are:

- **Hierarchical GNN Feature Extraction:** We leverage an hierarchical GNN pipeline to obtain three levels of representations—atom-level, motif-level, and graph-level—thereby preserving diverse topological signals essential for accurate property prediction.
- **Two-Level Hierarchical Mixture-of-Experts:** We introduce a low-level MoE that processes each granularity with specialized experts, along with a high-level MoE that integrates these outputs via topology-aware gating.
- **Multi-Scale Topology Segregation:** We extend topology-aware gating across multiple scales, ensuring that molecules are differentiated based on both local substructures and global properties.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

Molecular property prediction can be framed as a graph-based learning problem. In this setting, each molecule is represented as an attributed graph where atoms serve as vertices and chemical bonds as edges. Formally, let the training dataset be $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^{N}$, where each molecular graph $G_i \in \mathcal{G}$ is paired with a property vector $y_i \in \{0, 1\}^T$ that encodes T distinct characteristics. The goal is to learn a function $f : \mathcal{G} \to \mathcal{Y}$ that generalizes well to unseen molecular structures.

2.2 MOTIFS AND SCAFFOLDS

Molecular structures can be decomposed into two key components:

- **Scaffolds:** A molecular scaffold is the core structural framework of a molecule, primarily consisting of its ring systems and the linkers connecting them. This backbone remains constant while different functional groups or side chains can be modified to create diverse compounds. In drug discovery, scaffolds help researchers systematically explore chemical variations to optimize properties like potency, selectivity, and pharmacokinetics. The Bemis-Murcko framework (Bemis & Murcko (1996)) is a widely used method for defining scaffolds by identifying ring structures and linkers while disregarding terminal side chains.
- **Motifs:** Motifs are recurring structural patterns within a molecule that influences its chemical properties and reactivity. Often referred to as functional groups, they define how a molecule interacts in chemical reactions. The specific arrangement of these motifs within a molecule plays a key role in shaping its behavior and function(Fey et al. (2020); Peng et al. (2020)).

2.3 GRAPH NEURAL NETWORKS

Graph Neural Networks (GNNs, Gilmer et al. (2017) extract rich representations of molecular graphs through iterative message passing. At each layer l, the representation of an atom u is updated by aggregating information from its neighbors $\mathcal{N}(u)$ along with the associated bond features. This update is given by:

$$h_u^{(l+1)} = \Phi_u^{(l)}\left(h_u^{(l)},\ \Psi_a^{(l)}\left(\{(h_v^{(l)}, h_u^{(l)}, e_{uv}) : v \in \mathcal{N}(u)\}\right)\right), \tag{1}$$

where $\Phi_u^{(l)}$ and $\Psi_a^{(l)}$ are learnable functions for updating the node state and aggregating messages, respectively. A global molecular representation h_G is then obtained by applying a permutation-invariant readout function Ω over all atom embeddings:

$$h_G = \Omega\left(\{h_u^{(l)} \mid u \in G\}\right). \tag{2}$$

2.4 Mixture of Experts

The Mixture of Experts (MoE) framework Jacobs et al. (1991) leverages multiple specialized networks (experts) coordinated by a gating mechanism that routes each input to the most appropriate experts. Recent advancements in sparse MoE architectures Shazeer et al. (2017); Lepikhin et al. (2020); Du et al. (2022) have demonstrated that selectively activating a subset of experts can efficiently scale model capacity while managing computational costs. This approach has been widely adopted in large-scale language models Fedus et al. (2022a); Du et al. (2022); Fedus et al. (2022b), where MoE models often achieve performance comparable to or better than dense models with fewer resources.

3 Methods

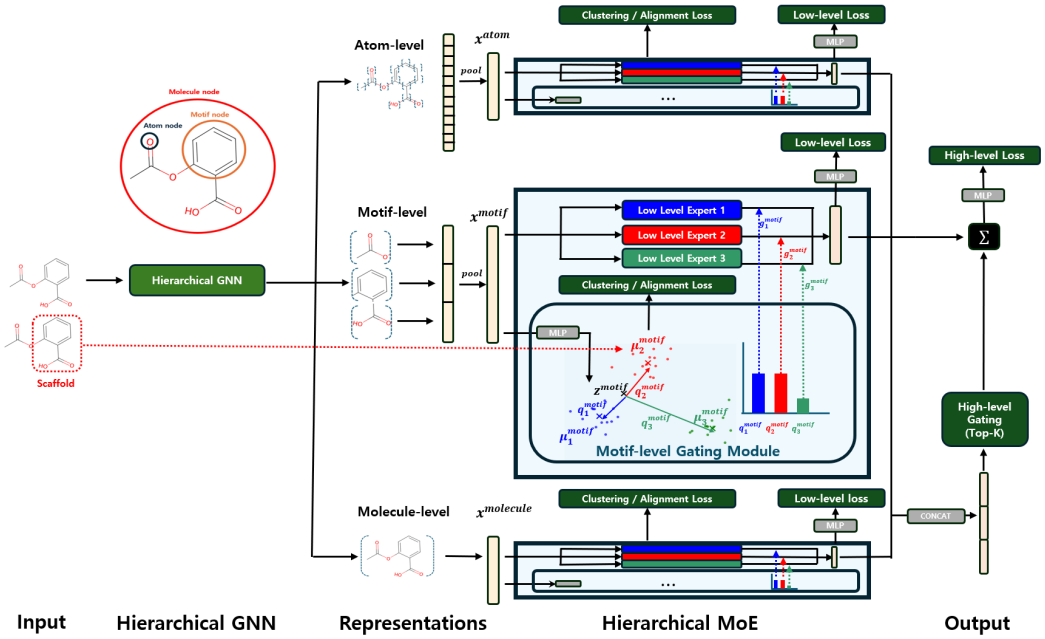

Figure 1: Overview of HierMolMoE: a hierarchical framework for molecular property prediction. Our model first uses a Hierarchical GNN—enhanced with motif and global nodes—to extract multi-scale representations. These features are then processed by a low-level Mixture-of-Experts module with topology-aware gating, and a high-level gating network integrates the outputs. During inference, gating modules select the most appropriate experts based on the molecule's scaffold.

3.1 Overview

Our framework addresses molecular property prediction through a hierarchical approach that explicitly models multiple structural granularities. The key insight is that different molecular properties may depend on features at varying scales - from local atom interactions to global molecular structure with the dependency on the overall topology of the molecule. Rather than forcing a single model to capture all these scales, we employ specialized experts at each level.

The framework consists of three main components:

1. **Multi-Scale Feature Extraction:** A hierarchical GNN processes the input molecular graph to generate representations at three distinct granularities: (1) atom-level, capturing local

chemical environments and bonding patterns, (2) motif-level, encoding functional groups and recurring substructures, and (3) graph-level, representing global molecular properties.

2. **Topology-Aware Expert Specialization:** Each granularity feeds into its own low-level mixture-of-experts module. These modules contain multiple expert networks that specialize in different chemical scaffolds or structural patterns. A topology-aware gating mechanism routes each input to the most relevant experts, allowing the model to develop specialized predictors for different classes of molecular structures.
3. **Dynamic Multi-Scale Integration:** A high-level mixture-of-experts module dynamically integrates predictions across all granularities. This module learns to weight the contributions of different scales based on the specific property being predicted and the input molecule's structure.

3.2 Model Architecture

3.2.1 Hierarchical Graph Neural Network

The foundation of our architecture is a hierarchical GNN that processes molecular graphs at multiple scales. Given an input molecular graph $G = (V, E)$, we enhance it with additional structural information by introducing motif nodes and a global graph node. For motif extraction, we follow the method introduced in Zhang et al. (2021), which uses BRICS decomposition (Degen et al. (2008)) to identify groups of atoms forming recurrent substructures. These substructures are then added as new motif nodes that connect to every atom in the corresponding motif, effectively augmenting the original graph with extra edges that encode local groupings.

For each node v, the GNN updates its representation through message passing as done in equation 1. After obtaining node embeddings though several layers of message passing, we partition them into three distinct sets:

- $H_{\text{atom}} \in \mathbb{R}^{n_a \times d}$: Atom-level representations
- $H_{\text{motif}} \in \mathbb{R}^{n_m \times d}$: Motif-level representations
- $H_{\text{graph}} \in \mathbb{R}^{1 \times d}$: Graph-level representation

Each set is then pooled to obtain fixed-size representations: x_{atom}, x_{motif}, and x_{graph} respectively.

3.2.2 Low-Level Mixture-of-Experts

For each granularity $g \in \{\text{atom}, \text{motif}, \text{graph}\}$, we employ a separate low-level MoE module with K_{low} experts. Each expert specializes in specific molecular substructures or patterns, allowing the model to capture different aspects of molecular topology at each granularity level. Following Kim et al. (2023), the low-level MoE process consists of several key steps:

Gating Mechanism In order to properly gate according to topology and granularity, we follow the method introduced in Kim et al. (2023). We first project each molecule's granularity-specific representation into a latent space. Next, soft expert assignments are computed using a Student's t-distribution, refined through Gumbel-Softmax sampling. Finally, these assignments are aligned with scaffold embeddings to incorporate prior chemical knowledge.

Given a granularity-specific representation x_g (atom, motif, or graph-level), we first transform it into a topology-aware representation through a non-linear dimension reduction network:

$$z_g = \text{MLP}(x_g), \tag{3}$$

Next, we maintain K_{low} learnable cluster centroids $\{C_k\}_{k=1}^{K_{\text{low}}}$, where each $C_k \in \mathbb{R}^{d_{z_g}}$. The assignment probability to the k-th expert is computed using a Student's t-distribution with one degree of freedom:

$$q_k = \frac{(1 + \|z_g - C_k\|^2)^{-1}}{\sum_{k'=1}^{K_{\text{low}}} (1 + \|z_g - C_{k'}\|^2)^{-1}}, \tag{4}$$

This formulation naturally groups molecules with similar topological patterns to the same expert through a soft clustering mechanism. To enable stochastic expert selection during training while maintaining differentiability, we apply Gumbel-Softmax:

$$g_k = \frac{\exp((\log q_k + \gamma_k)/\tau)}{\sum_{k'=1}^{K_{\text{low}}} \exp((\log q_{k'} + \gamma_{k'})/\tau)}, \tag{5}$$

where γ_k is drawn from a Gumbel distribution, and τ is the temperature parameter annealed from a high initial value τ_0 to a low final value τ_E during training. As training progresses and τ decreases, the gating weights approach a one-hot distribution, and the annealing process gradually transitions from exploring multiple experts to specializing in specific molecular topologies.

To additionally incorporate prior knowledge of molecular topology, we additionally align the expert assignments with molecular scaffolds. For a training set with $|S|$ scaffolds, each molecule's scaffold index is represented as a one-hot vector. We maintain learnable scaffold embeddings $\{\varepsilon_s\}_{s=1}^{|S|}$ where $\varepsilon_s \in \mathbb{R}^{d_{z_g}}$, and define a cost matrix $M \in \mathbb{R}^{|S| \times K_{\text{low}}}$ based on cosine distances between scaffold embeddings and cluster centers:

$$m_{sk} = 1 - \cos(\varepsilon_s, C_k). \tag{6}$$

While we also experimented with molecular fingerprint-based clustering, we found that scaffold-based alignment consistently yields better performance in practice.

Expert Integration. Each expert f_k is implemented as a single fully-connected layer and independently processes the input to produce task-specific predictions. The outputs are combined using the gating weights:

$$y_g = \sum_{k=1}^{K_{\text{low}}} g_k f_k(x_g), \tag{7}$$

producing granularity-specific predictions $y_g \in \mathbb{R}^T$ for T tasks.

3.2.3 High-Level Integration

The high-level MoE module combines predictions from all granularities. First, we concatenate the topology-aware latent representations z_g that were used for gating in the low-level MoE:

$$L = [z_{\text{atom}}; z_{\text{motif}}; z_{\text{graph}}], \tag{8}$$

which is processed by a gating network to compute weights for each granularity:

$$w = \text{softmax}(W_h L + b_h). \tag{9}$$

The final prediction is computed as a weighted combination:

$$y_{\text{final}} = \sum_{g \in \{\text{atom,motif,graph}\}} w_g y_g. \tag{10}$$

3.3 Training and Loss Functions

3.3.1 Training Objective

The overall training objective combines three types of losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \alpha \mathcal{L}_{\text{cluster}} + \beta \mathcal{L}_{\text{align}}, \tag{11}$$

where α and β are balancing parameters, $\mathcal{L}_{\text{pred}}$ is the primary prediction loss, $\mathcal{L}_{\text{cluster}}$ is the clustering loss for encouraging cohesive expert specialization, and $\mathcal{L}_{\text{align}}$ is the scaffold alignment loss.

3.3.2 TRAINING STRATEGY

To ensure stable training and effective expert specialization, we adopt a two-stage training process:

1. **Warmup Phase:** Each low-level MoE is first trained independently with a high temperature τ, allowing experts to explore diverse molecular patterns.
2. **Joint Training Phase:** The entire network is trained end-to-end while gradually annealing τ from τ_0 to τ_E. We alternate between epochs focusing on joint prediction loss and those emphasizing individual expert specialization.

During inference, we directly use the cluster assignment probabilities q_k instead of the Gumbel-Softmax outputs for deterministic prediction.

3.3.3 LOSS FUNCTIONS

For molecular property prediction tasks, we use binary cross-entropy as our prediction loss:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \text{BCE}(y_i, \hat{y}_i), \tag{12}$$

where y_i is the ground truth label and $\hat{y}_i$ is the model's prediction. The selection of $\hat{y}_i$ differs according to the training phase.

To strengthen cluster cohesion, we define a clustering loss using a target distribution that sharpens the assignments:

$$p_k = \frac{q_k^2 / \sum_i q_{i,k}}{\sum_{k'} (q_{k'}^2 / \sum_i q_{i,k'})}, \tag{13}$$

$$\mathcal{L}_{\text{cluster}} = \text{KL}(P \| Q) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K_{\text{low}}} p_{i,k} \log \frac{p_{i,k}}{q_{i,k}}. \tag{14}$$

For the scaffold alignment, we follow Kim et al. (2023) that encourages consistency between expert assignments and molecular scaffolds:

$$\mathcal{L}_{\text{align}} = \sum_{s=1}^{|S|}\sum_{k=1}^{K_{\text{low}}} t_s \cdot q_k \cdot m_{sk}, \tag{15}$$

where t_s and q_k are the scaffold and cluster assignment probabilities respectively and $m_s k$ is defined in equation 6. This alignment loss ensures molecules sharing the same scaffold are assigned to similar experts, while allowing topologically similar scaffolds to be grouped together when $|S| \gg K_{\text{low}}$.

4 EXPERIMENTS

We conduct extensive experiments to evaluate HierMolMoE on multiple molecular property prediction benchmarks. First, we compare our method against state-of-the-art baselines across various molecular prediction tasks. Then, we perform detailed ablation studies to analyze the contribution of each hierarchical component.

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS

We evaluate our approach on eight benchmark datasets widely employed for molecular property prediction Wu et al. (2018). These datasets span diverse molecular prediction tasks - from toxicity

prediction (Tox21, ToxCast, SIDER) to pharmacological properties (BBBP, ClinTox) - providing a comprehensive testbed for assessing model performance. Following standard practice Kim et al. (2023); Hu et al. (2021), we extract molecular features including atom attributes, bond characteristics, and scaffold indices using the RDKit toolkit Landrum (2013).

To ensure rigorous evaluation of generalization capability, we adopt the scaffold splitting protocol Hu et al. (2021) which partitions molecules based on their structural scaffolds into training, validation, and test sets (80:10:10 ratio). This protocol is more challenging than random splitting as it requires models to predict properties of molecules with entirely novel scaffolds not seen during training.

4.1.2 Implementation Details

We train all models for a maximum of 500 epochs with early stopping if validation AUROC does not improve for 50 consecutive epochs. For HierMolMoE, we employ 4 experts per granularity, resulting in 12 low-level experts total. Model parameters are optimized using Adam with a learning rate of $1e-4$. All runs were run in a single NVIDIA GeForce RTX 3090 TURBO GDDR6X 24GB.

Our two-stage training process begins with a 50-epoch warmup phase where granularity experts are trained independently. The validation monitoring starts after this warmup during joint training. For Gumbel-Softmax sampling, we employ temperature annealing from $\tau_0 = 10$ to $\tau_E = 0.01$. For clustering and alignment loss weights, we fix $\alpha = 0.1$ and $\beta = 0.01$. All experiments use GIN Xu et al. (2018) as the backbone with identical node encoders. Results are reported as mean AUROC (±standard deviation) across 10 different random seeds.

4.1.3 Baselines

We compare against the following several methods using identical GIN backbones but differing in their prediction strategies:

- **Single Classifier (SingleCLF)** Xu et al. (2018): A standard classifier that employs a single prediction module (i.e., one expert) for generating the final output.
- **Mixture of Experts (MoE)** Zoph et al. (2022): A model that utilizes an MLP with Gumbel-Softmax to stochastically select and combine the outputs from multiple experts.
- **Expert-Ensemble (E-Ensemble)** Dietterich (2000): An approach that aggregates the outputs from several experts by taking their arithmetic mean.
- **GraphDIVE** Hu et al. (2021): A method that combines expert outputs using a weighted sum, where the weights are computed by a linear layer followed by a Softmax.
- **MoCE:** Yao et al. (2023) A method that combines a GNN encoder with a dynamic multi-expert predictor, where a gating network selectively weights diverse expert projections to capture both common and topology-specific molecular features for property prediction
- **TopExpert** Kim et al. (2023): A topology-specific MoE that leverages a clustering-based gating module to assign molecules into groups according to their topological features.

4.2 Results and Discussion

4.2.1 Main Results

Table 1 presents the comparison between HierMolMoE and state-of-the-art baselines across eight benchmark datasets. Our method consistently outperforms all baselines, achieving an average AUC-ROC of 74.1% and representing a significant improvement of 5.5% over the next best baseline (TopExpert, 68.0%). The improvements are particularly pronounced on ClinTox (80.3% vs 62.0%), BACE (82.6% vs 71.4%), and MUV (79.5% vs 72.9%). Moreover, our method demonstrates remarkable stability with lower standard deviations across most datasets - notably BBBP (±0.6 vs ±3.5).

Model	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	AVG
# of Mol.	2,039	7,831	8,575	1,427	1,478	93,087	41,127	1,513	-
# of Tasks	1	12	617	27	2	17	1	1	-
SingleCLF	69.9 ± 3.1	73.9 ± 1.4	64.1 ± 1.2	58.7 ± 1.5	60.1 ± 3.7	74.1 ± 3.3	73.3 ± 2.8	67.7 ± 3.9	67.8
MoE	62.8 ± 4.1	74.6 ± 0.3	59.1 ± 1.4	58.5 ± 1.9	57.5 ± 2.1	76.9 ± 0.7	73.4 ± 1.5	69.1 ± 2.1	66.6
E-Ensemble	67.1 ± 1.3	74.1 ± 0.8	60.3 ± 0.9	55.9 ± 1.3	60.1 ± 4.6	71.9 ± 3.7	76.2 ± 1.3	67.9 ± 3.9	66.7
GraphDIVE	65.2 ± 1.7	71.6 ± 2.6	57.7 ± 0.3	54.2 ± 3.7	59.3 ± 3.9	69.1 ± 4.0	70.1 ± 1.1	62.3 ± 4.1	63.7
MoCE	64.6 ± 2.5	72.0 ± 1.2	62.0 ± 0.6	57.4 ± 1.0	57.3 ± 1.0	67.3 ± 3.9	70.9 ± 4.2	71.4 ± 2.8	65.4
TopExpert	69.5 ± 3.5	73.9 ± 0.7	61.3 ± 0.6	56.8 ± 1.3	62.0 ± 2.3	72.9 ± 5.0	76.6 ± 1.9	71.3 ± 4.6	68.0
Ours	**71.9** ± 0.6	**76.3** ± 0.4	**64.7** ± 0.5	**59.4** ± 1.6	**80.3** ± 3.9	**79.5** ± 1.0	**78.1** ± 0.5	**82.6** ± 1.7	**74.1**

Table 1: Performance comparison((↑) with baseline models on molecular property prediction tasks. Values indicate the mean AUC-ROC (%) across datasets with **best values** in bold.

	Model Components			Performance Metrics (↑)							Rank (↓)
	Atom	Motif	Graph	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	AVG	
Single MoE	✓			61.2	75.9	63.7	59.4	65.5	76.6	67.0	5.86
		✓		59.5	72.3	64.8	56.4	66.0	81.0	66.7	6.29
			✓	56.7	75.2	**66.5**	60.8	69.5	81.4	68.4	4.29
Dual MoE	✓	✓		<u>70.6</u>	75.4	64.4	60.3	**82.8**	82.0	<u>72.6</u>	3.29
	✓		✓	69.6	**76.6**	66.0	**61.7**	77.0	<u>82.7</u>	72.3	<u>2.29</u>
		✓	✓	68.4	73.4	<u>66.4</u>	60.5	75.8	81.2	71.0	4.14
Triple MoE	✓	✓	✓	**71.4**	<u>76.5</u>	65.2	<u>60.9</u>	<u>81.3</u>	**83.4**	**73.1**	**1.86**

Table 2: Performance comparison(↑) across different granularity combinations. Each row represents a different model configuration using atom, motif, and graph-level experts. Results show mean AUC-ROC (%) for 5 random seed runs with **best** and <u>second-best</u> values in bold and underlined respectively. The Rank column shows the average ranking across all metrics (lower is better).

4.3 Discussion

Table 1 presents the comparison between HierMolMoE and baseline approaches across eight benchmark datasets. Our method achieves superior overall performance with an average AUC-ROC of 74.1%, a 5.7% improvement over the next best baseline. The improvements are particularly significant on ClinTox (79.3%), BACE (81.6%), and MUV (79.5%). Beyond better accuracy, our method shows notably lower standard deviations, indicating more stable predictions. These results demonstrate that modeling molecular properties through granularity-specific experts offers a more effective approach than using a single GNN model, highlighting the importance of specialized representations at different structural scales.

To further understand the effectiveness of our hierarchical design, we conduct detailed ablation studies by systematically varying model components (Table 2). In single granularity settings, we observe that different datasets favor different granularities. Notably, models using multiple granularities (dual or triple) consistently outperform single granularity variants across all datasets, with improvements of up to 20% (ClinTox: 65.5% $\rightarrow$ 82.8%). While the Triple MoE configuration is not always the top performer for every individual dataset, it exhibits overall robust and balanced performance compared to both the dual and single granularity models. This indicates that integrating all three levels of representation not only provides competitive predictions on a per-task basis but also offers a more consistent and generalizable approach across diverse molecular property prediction tasks.

Overall, our findings suggest that the conventional approach of using a single granularity may be limiting the ability to fully capture molecular properties. By explicitly modeling and combining multiple structural scales through topology-aware specialized experts, we can better handle the inherent complexity of molecular structures and their associated properties.

Meaningfulness Statement

All systems of life possess inherent hierarchical organization, from atoms to molecules to complex biological systems. This multi-scale organization is fundamental to how biological systems process information and determine their properties. Our work contributes to this direction by developing a robust molecular representation framework that mirrors nature's hierarchical structure, enabling more reliable predictions of molecular properties through specialized experts at different structural levels.

ACKNOWLEDGMENTS

I'd like to thank Sanghoon Lee, Jungwoo Park and Junhyun Lee for the helpful feedbacks.

This work was supported in part by the Ministry of Science and ICT / Ministry of Health and Welfare [RS-2024-00523644]; Ministry of SMEs and Startups [RS-2024-00523644]; and the Ministry of Health and Welfare [RS-2020-KH088565].

REFERENCES

Jinheon Baek, Minki Kang, and Sung Ju Hwang. Accurate learning of graph representations with graph multiset pooling. *arXiv preprint arXiv:2102.11533*, 2021.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Yuzhou Chen and Yulia R Gel. Topological pooling on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7096–7103, 2023.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning*, pp. 5547–5569, 2022.

William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022a.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022b.

Matthias Fey, Jan-Gin Yuen, and Frank Weichert. Hierarchical inter-message passing for learning on molecular graphs. *arXiv preprint arXiv:2006.12179*, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Fenyu Hu, Liping Wang, Shu Wu, Liang Wang, and Tieniu Tan. Graph classification by mixture of diverse experts. *arXiv preprint arXiv:2103.15622*, 2021.

Muhammad Ifte Khairul Islam, Max Khanov, and Esra Akbas. Mpool: motif-based graph pooling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 105–117. Springer, 2023.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Hossam Kadry, Behnam Noorani, and Luca Cucullo. A blood–brain barrier overview on structure, function, impairment, and biomarkers of integrity. *Fluids and Barriers of the CNS*, 17:1–24, 2020.

Suyeon Kim, Dongha Lee, SeongKu Kang, Seonghyeon Lee, and Hwanjo Yu. Learning topology-specific experts for molecular property prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8291–8299, 2023.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Hao Peng, Jianxin Li, Qiran Gong, Yuanxin Ning, Senzhang Wang, and Lifang He. Motif-matching based subgraph-level attentional convolutional network for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5387–5394, 2020.

Kunal Roy and Partha Pratim Roy. Qsar of cytochrome inhibitors. *Expert Opinion on Drug Metabolism & Toxicology*, 5(10):1245–1266, 2009.

T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Jie Shen and Christos A Nicolaou. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36, 2019.

Eduardo Soares, Indra Priyadarsini, Emilio Vital Brazil, Victor Yukio Shirasuna, and Seiji Takeda. Multi-view mixture-of-experts for predicting molecular properties using smiles, selfies, and graph-based representations. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024.

Fang Wu, Dragomir Radev, and Stan Z Li. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5312–5320, 2023.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

Xu Yao, Shuang Liang, Songqiao Han, and Hailiang Huang. Enhancing molecular property prediction via mixture of collaborative experts. *arXiv preprint arXiv:2312.03292*, 2023.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.