

BOOSTING FEW-SHOT LEARNING IN DRUG DISCOVERY VIA MOLECULAR HIERARCHICAL MATCHING

Anonymous authors

Paper under double-blind review

ABSTRACT

Drug discovery is essential for identifying candidate drugs for various diseases. However, its low success rate often results in a scarcity of annotations, leading to the few-shot learning problem. Existing approaches primarily focus on single-scale structural features, which may overlook crucial hierarchical information within molecular structures. Considering that different structural levels determine different molecular properties, we introduce **Hierarchical Matching Networks (HierMatch)**. This approach utilizes hierarchical pooling and matching techniques to enhance the representation and comparison of molecular features across multiple scales. Specifically, hierarchical pooling aggregates node features from local neighborhoods to global structures, preserving essential details that single-scale methods may overlook. Hierarchical matching then performs pairwise comparisons at these different structural levels, allowing more precise assessments of molecular similarities. *To further enhance our model, we incorporate a meta-learning strategy that independently updates general and task-specific parameters, improving feature alignment and matching across hierarchical levels.* This allows our approach adaptively captures subtle structural similarities and distinctions, improving both predictive accuracy and generalization. Our experimental results show that HierMatch outperforms state-of-the-art (SOTA) methods on the MoleculeNet and FS-Mol benchmarks, with improvements of 2.87% in AUROC and 6.52% in Δ AUPRC. Additionally, HierMatch demonstrates excellent generalization performance on the Meta-MolNet benchmark.

1 INTRODUCTION

Drug discovery is pivotal for human health, involving the screening and optimization of numerous compounds to identify potential drug candidates that satisfy both pharmacological efficacy and toxicological safety criteria (Drews, 2000; Renaud et al., 2016; Atanasov et al., 2021). The traditional drug development cycle typically spans over a decade, incurs costs exceeding 1 billion dollars, yet achieves a success rate of less than 10% (Sliwoski et al., 2014; Adelusi et al., 2022). Artificial Intelligence-Driven Drug Discovery (AIDD) has emerged as a promising solution to address this challenge (Mak et al., 2023; Macalino et al., 2015; Gawehn et al., 2016). Within AIDD, Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) (Cherkasov et al., 2014; Liu & Long, 2009) models are crucial for predicting the relationships between molecular structures and their activities. These methods (Zhang et al., 2021a; Fabian et al., 2020; Wang et al., 2021; Chen et al., 2023) rely heavily on extensive datasets due to the complexity of understanding and modeling molecular geometries. However, the lengthy durations, high costs, and low success rates of chemical wet experiments limit the availability of experimental data, resulting in a scarcity of labeled data.

Few-shot learning (Li et al., 2023; Wang et al., 2020b) has shown potential in addressing this data scarcity, enabling models to generalize quickly from minimal data to new tasks. Most approaches are based on molecular graphs with atoms as nodes and chemical bonds as edges, using Graph Neural Networks (GNNs) (Zhou et al., 2020) to capture molecular topological structures. In particular, IterRefLSTM (Altae-Tran et al., 2017), Meta-MGNN (Guo et al., 2021), PAR (Wang et al., 2021), ADKF-IFT (Chen et al., 2023), and Meta-GAT (Lv et al., 2024) use GNNs as encoders to learn molecular representations for label inference. *Conversely, several sequence-based methods, such as CHEF* (Adler et al., 2020) and MHNfs, (Schimunek et al., 2023) utilize Multilayer Perceptrons (MLPs) as encoders to compress molecular fingerprints or descriptors for predictive modeling.


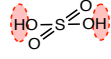
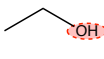

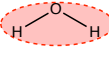
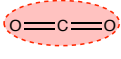
Atom			Substructure		Molecule			
Hydrogen chloride		Sulfuric acid	Ethanol	Dodecane	Water	Carbon dioxide		
								
1	Acidity	1	0	Hydrophobicity	1	1	Boiling point	0
The hydro atoms affect the molecular acidity.			The hydroxyl groups affect the molecular hydrophobicity.		The whole structures affect the molecular boiling point.			
(a)			(b)		(c)			

Figure 1: Different structures of molecules affect different molecular properties. (a) At the atom level, hydrogen ions in hydrochloric and sulfuric acids determine acidity. (b) At the substructure level, hydroxyl groups in ethanol and dodecane affect hydrophobicity. (c) At the molecule level, the overall structures influence boiling points. Red blocks highlight key molecular structures.

However, existing approaches often overlook a crucial aspect: **different levels of structural information—ranging from atoms to substructures and the entire molecule—determine distinct molecular properties**. Some properties are influenced by atomic composition, while others depend on substructures or the overall molecular configuration. Figure 1 provides corresponding examples: (a) shows how hydrogen ions determine the acidity of hydrochloric acid and sulfuric acid; (b) demonstrates how hydroxyl groups influence the hydrophobic properties of ethanol and dodecane; (c) indicates how the overall molecular structure affects boiling points. In graph-based methods, using multiple GNN layers may cause over-smoothing, where the receptive fields of nodes expand to cover the entire molecular graph, thus obscuring substructural details. This makes GNNs more suitable for predicting properties related to the overall structure of molecules. Conversely, fingerprint-based methods offer only fragmented local features, potentially overlooking critical information from the overall molecular structure. Although CHEF (Adler et al., 2020) introduces a representation fusion strategy, its reliance on ECFP6 (Rogers & Hahn, 2010)—which is based on fixed local features—limits its capacity for multi-level structural analysis. Therefore, capturing different levels of molecular structures is crucial for accurately predicting various molecular properties.

To address this challenge, we propose **Hierarchical Matching Networks (HierMatch)**, which utilizes hierarchical pooling and matching to enhance the capabilities of GNNs for few-shot molecular representation learning. This approach enables the model to more accurately capture and predict molecular properties in few-shot scenarios. Our main contributions are summarized as follows:

- To the best of our knowledge, we are pioneers in applying a hierarchical structural matching approach to few-shot learning for molecular property prediction. This approach enhances the model’s understanding of complex molecular structures via hierarchical pooling and matching, seamlessly integrating with existing GNNs while significantly boosting performance with minimal additional cost (Section 3).
- We propose a hierarchical framework that integrates information from atoms to higher-level structures, capturing complex molecular features. Using an attention-based matching module, the model aligns representations at various levels, selecting the most relevant features for improved prediction (Sections 3.3.1 and 3.3.1).
- We propose an adaptive meta-learning strategy that independently updates task-specific and general parameters while dynamically fine-tuning parameters across multiple hierarchical levels, further enhancing the effectiveness of hierarchical matching (Section 3.3.2).
- Our HierMatch outperforms the SOTAs on both the MoleculeNet (Section 4.1) and FS-Mol (Section 4.2) benchmarks, achieving improvements of 2.87% in AUROC and 6.52% in Δ AUPRC, respectively. Additionally, we test the generalization ability of HierMatch on the Meta-Mol benchmark, which shows outstanding performance (Section 4.3).

2 RELATED WORK

2.1 GRAPH-BASED MOLECULAR PROPERTY PREDICTION

Molecular graphs, where atoms as nodes and chemical bonds as edges, represent the complex topology of molecules, facilitating a deeper understanding of their physical and chemical properties. Graph Neural Networks (GNNs) have been employed to encode these structural features, with notable examples including Graph Isomorphism Networks (GIN) (Xu et al., 2019), Graph Convolutional Networks (GCN) (Kipf & Welling, 2017), Graph Attention Networks (GAT) (Veličković et al., 2018), and Message Passing Neural Networks (MPNN) (Gilmer et al., 2017).

Graph-based methods are also mainstream for the few-shot molecular property prediction task. PAR (Wang et al., 2021) and ADKF-IFT (Chen et al., 2023) employ GIN (Xu et al., 2019) as the molecular encoder, while Meta-MGNN (Guo et al., 2021) utilizes Pre-GIN (Hu et al., 2020). Meta-GAT (Lv et al., 2024) adopts GAT (Veličković et al., 2018) to learn molecular representations. However, these methods (Altae-Tran et al., 2017; Ren et al., 2018; Zhuang et al., 2023; Chen et al., 2023; Schimunek et al., 2023) overlook the impact of varying molecular structures on different molecular properties or activities. In addition, several approaches (Zhao et al., 2023; Liu et al., 2024) combine the strengths of LLMs to tackle the few-shot problem, but these methods incur high computational costs. Our method differs by incorporating molecular hierarchical structures through hierarchical pooling and matching, allowing for more effective representation of complex structures.

2.2 MATCHING LEARNING

In few-shot learning, matching learning compares new instances with a small set of labeled examples to facilitate accurate predictions, with methods like Matching Networks (Vinyals et al., 2016), ProtoNet (Snell et al., 2017), Relation Networks (Sung et al., 2018), and LGM-Net (Li et al., 2019). Hierarchical matching further enhances few-shot learning by matching multi-level representations. Specifically, AMN (Mai et al., 2019) enhances embedding robustness with feature-level attention mechanisms. Advancedly, SSF-HRNet (Zhong et al., 2023) incorporates self-similarity features and hierarchical relationships to boost performance. Additionally, VTM (Kim et al., 2023) employs a hierarchical encoder-decoder architecture to perform patch-level non-parametric matching. Similarly, HCL (Zheng et al., 2022) combines hierarchical matching with contrastive learning to identify discriminative patches. These methods have been applied in Natural Language Processing (NLP) and Computer Vision (CV). However, related research remains sparse in drug discovery. To address this gap, we propose HierMatch, which leverages structural hierarchical matching specifically for the few-shot molecular property prediction task.

2.3 HIERARCHICAL REPRESENTATION LEARNING ON GRAPHS

Hierarchical representation learning is vital for graphs as it captures multi-scale structures, allowing models to discern both local and global patterns more effectively (Grattarola et al., 2022). Existing methods in this domain can be divided into three categories: layer-wise hierarchical methods, architecture-level hierarchical methods, and supergraph-based methods. Layer-wise hierarchical methods, such as DiffPool (Ying et al., 2018), JK-Nets (Xu et al., 2018), Top-K Pooling (Lee et al., 2019), ASAP (Ranjan et al., 2020), and HGP-SL (Zhang et al., 2019), pool nodes into clusters to create coarser graph representations. In contrast, MixHop (Abu-El-Haija et al., 2019) captures higher-order graph structures by leveraging multiple powers of the adjacency matrix for feature aggregation. Additionally, architecture-level hierarchical methods like FraGAT (Zhang et al., 2021b) and MGSSL (Zhang et al., 2021a) employ branched networks for molecular multi-scale representation learning at atom, structure, and molecule levels, while HiMol (Zang et al., 2023) adopts supergraph construction to represent hierarchical information. Unlike these methods, our HierMatch not only obtains multi-level molecular representations but also utilizes matching learning to capture properties influenced by different hierarchical structures.

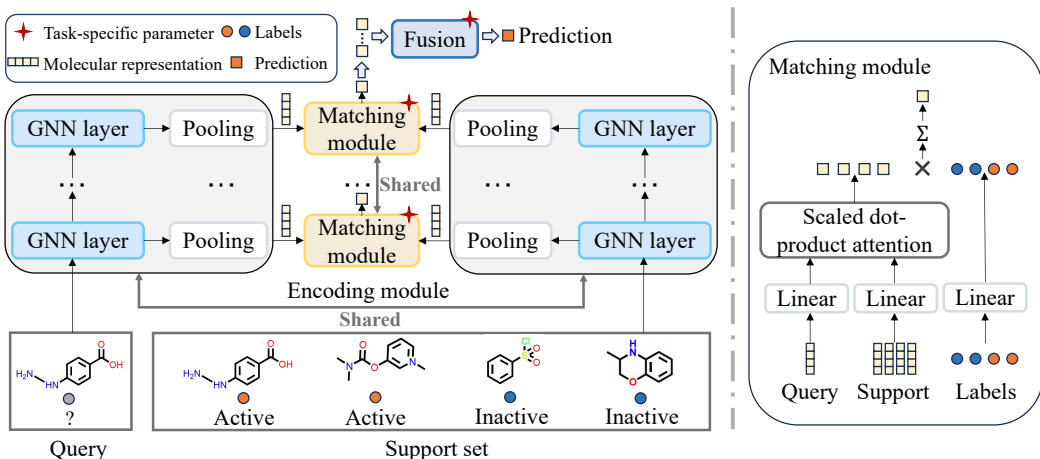


Figure 2: The overview of HierMatch. **Left:** Our model follows a hierarchical pooling-matching architecture comprising two components: an encoding module (including pooling) and a matching module. First, mean pooling is applied at each GNN layer to generate multi-level molecular representations. Then, an attention mechanism is utilized to align representations between the support set and query set across different levels. Finally, predictions from different GNN layers are integrated to obtain the final results. **Right:** The detailed process of the matching module.

3 METHOD

We propose **Hierarchical Matching Networks (HierMatch)**, a simple yet effective approach for the few-shot molecular property prediction task. Figure 2 provides an overview of the HierMatch framework. In this section, we first introduce the few-shot molecular property prediction problem in Section 3.1. Then, we discuss the background of GNNs in Section 3.2. Finally, we detail our method and its training process in Section 3.3.

3.1 PROBLEM DEFINITION

The few-shot molecular property prediction problem, as defined by ADKF-IFT (Chen et al., 2023) and MHNfs (Schimunek et al., 2023), involves training models on a set of tasks $\{\mathcal{T}_\tau\}_{\tau=1}^{N_t}$ from the training set \mathcal{D}_{train} to enhance generalization to new tasks. Each task \mathcal{T}_τ includes a support set $\mathcal{S}_\tau = \{(\mathbf{x}_{\tau,i}, y_{\tau,i})\}_{i=1}^{N_\tau^s}$ and a query set $\mathcal{Q}_\tau = \{(\mathbf{x}_{\tau,j}, y_{\tau,j})\}_{j=1}^{N_\tau^q}$, where $\mathbf{x}_{\tau,i} \in \mathbb{R}^d$ and $\mathbf{x}_{\tau,j} \in \mathbb{R}^d$ represent molecular features, and $y_{\tau,i}, y_{\tau,j} \in \{0, 1\}$ indicate the molecular properties or activities. The support set \mathcal{S}_τ provides a few labeled examples for task-specific adaptation, while the query set \mathcal{Q}_τ is utilized to evaluate the model’s performance on unseen examples.

3.2 PRELIMINARIES

Graph neural networks (GNNs) are designed to handle graph-structured data (non-Euclidean data), by aggregating information from neighboring nodes to learn effective representations (Zhou et al., 2020). Models such as GCN (Kipf & Welling, 2017), GIN (Xu et al., 2019), and GAT (Veličković et al., 2018) are widely used for tasks like graph classification and other related applications. In a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, \mathcal{V} represents the set of nodes and \mathcal{E} the set of edges. $\mathbf{h}_v^{(0)}$ represents the initial features of node v , and $\mathbf{b}_{u,v}$ denotes the features of the edge $e_{u,v}$ between nodes u and v . At the l^{th} layer, the representation $\mathbf{h}_v^{(l)}$ of node v is updated in GNNs as follows:

$$\mathbf{h}_v^{(l)} = \text{UPDATE}^{(l)} \left(\mathbf{h}_v^{(l-1)}, \text{AGGREGATE}^{(l)} \left(\left\{ \left(\mathbf{h}_u^{(l-1)}, \mathbf{b}_{u,v} \right) \mid u \in \mathcal{N}(v) \right\} \right) \right), \quad (1)$$

where $\mathcal{N}(v)$ is the set of neighboring nodes of v . The AGGREGATE function combines features from neighboring nodes, and the UPDATE function updates the node features for the next layer.

3.3 HIERARCHICAL MATCHING

Figure 1 demonstrates the critical role of hierarchical molecular structures in determining molecular properties. Consequently, predicting molecular properties from limited samples requires considering these hierarchical structures. As shown in Figure 2, we apply a pooling operation at each layer to extract multi-level molecular representations. Then, we employ hierarchical matching, utilizing an attention mechanism to align these representations and make predictions at hierarchical levels. The final prediction is obtained by integrating the outputs from various GNN layers.

3.3.1 MODEL ARCHITECTURE

Encoding Module. In Section 2.1, we highlight the importance of hierarchical molecular structures in determining molecular properties. Following the mainstream graph-based few-shot molecular property prediction approaches (Wang et al., 2021; Guo et al., 2021; Lv et al., 2024; Chen et al., 2023), we adopt the widely used GIN (Xu et al., 2019) as the backbone of our method. In GNNs, each layer aggregates local information from nodes and their neighboring hops. As the network depth increases, the model incrementally aggregates hierarchical information, from individual nodes to substructures, and ultimately to the entire molecule.

To capture molecular representations at different levels, we employ mean pooling to aggregate node representations at each layer of the GNN. For a given task τ , we first apply Eq. 1 to obtain the node representations $\mathbf{h}_{\tau,s}^{(l)} \in \mathbb{R}^{n_s \times d}$ for the support set \mathcal{S}_τ and $\mathbf{h}_{\tau,q}^{(l)} \in \mathbb{R}^{n_q \times d}$ for the query set \mathcal{Q}_τ . Then, we utilize mean pooling to derive the molecular representations $\mathbf{z}_{\tau,s}^{(l)} \in \mathbb{R}^{N_\tau^s \times d}$ for the support set \mathcal{S}_τ and $\mathbf{z}_{\tau,q}^{(l)} \in \mathbb{R}^{N_\tau^q \times d}$ for the query set \mathcal{Q}_τ , as follows:

$$\mathbf{z}_{\tau,s}^{(l)} = \text{Pooling}(\mathbf{h}_{\tau,s,v}^{(l)}, v \in \mathcal{V}_{\tau,s}), \mathbf{z}_{\tau,q}^{(l)} = \text{Pooling}(\mathbf{h}_{\tau,q,v}^{(l)}, v \in \mathcal{V}_{\tau,q}), \quad (2)$$

where the Pooling function denotes mean pooling, and l refers to the l^{th} layer of the GNN.

Matching Module. In few-shot learning, matching learning evaluates the similarity between a query sample and those in the support set (Vinyals et al., 2016; Wang et al., 2020a). We implement hierarchical matching in GNNs, focusing on molecular structures at different layers to improve the precision of similarity identification among molecules.

We adopt the attention mechanism introduced by Vaswani et al. (2017) to measure similarity. Specifically, we designate the molecular representations $\mathbf{z}_{\tau,s}^{(l)} \in \mathbb{R}^{N_\tau^s \times d}$ in the support set \mathcal{S}_τ as the key, and the molecular representations $\mathbf{z}_{\tau,q}^{(l)} \in \mathbb{R}^{N_\tau^q \times d}$ in the query set \mathcal{Q}_τ as the query. The corresponding ground-truth labels $\mathbf{y}_{\tau,s} \in \mathbb{R}^{N_\tau^s \times 1}$ in the support set are used as the value. This approach enables us to achieve matching results at the specified level using the attention mechanism:

$$\hat{\mathbf{y}}_{\tau,q}^{(l)} = \text{Softmax}\left(\frac{(\mathbf{z}_{\tau,q}^{(l)} \mathbf{W}_q)(\mathbf{z}_{\tau,s}^{(l)} \mathbf{W}_k)^\top}{\sqrt{d}}\right) \mathbf{y}_{\tau,s}, \quad (3)$$

where d is the dimension of molecular representations, and $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d \times d}$.

Fusion. We repeat the above steps to obtain the matching results $\hat{\mathbf{y}}_{\tau,q}^{(l)}$ for each GNN layer. These results are then concatenated to form a comprehensive representation, integrating predictions from different layers. This concatenated representation is subsequently passed through a Linear layer to produce the final prediction $\hat{\mathbf{y}}_{\tau,q} \in \mathbb{R}^{N_\tau^q \times 1}$:

$$\hat{\mathbf{y}}_{\tau,q} = \text{Linear}_{\mathbf{W}_o}(\text{Concat}(\hat{\mathbf{y}}_{\tau,q}^{(1)}, \hat{\mathbf{y}}_{\tau,q}^{(2)}, \dots, \hat{\mathbf{y}}_{\tau,q}^{(L)})), \quad (4)$$

where L represents the total number of layers in the GNN, and $\mathbf{W}_o \in \mathbb{R}^{L \times 2}$ are the parameters of the Linear function. This fusion process ensures that the final prediction leverages the hierarchical information captured at each layer, resulting in a more robust and comprehensive output. By effectively integrating multi-level features, the model better captures the complex relationships inherent in molecular structures.

¹the subscript s represents it belongs to support set.

²the subscript q represents it belongs to query set.

Algorithm 1 Meta-training procedure for HierMatch.**Input:** The few-shot training tasks $\{\mathcal{T}_\tau\}_{\tau=1}^{N_t}$ of molecular property prediction, learning rate η ,**Output:** trained model $f_{\theta, \mathbf{w}}$

```

1: Randomly initialize  $\theta$  and  $\mathbf{w}$ ;
2: while not converged do
3:   Sample a batch  $\mathcal{B}$  of tasks  $\mathcal{T}_\tau$ ;
4:   for all  $\mathcal{T}_\tau$  do
5:     Sample  $N_\tau^s$  and  $N_\tau^q$  molecules to form  $\mathcal{S}_{\text{train}, \tau}$  and  $\mathcal{Q}_{\text{train}, \tau}$ ;
6:     for  $l = 1, \dots, L$  do
7:       Obtain node representations  $\mathbf{h}_{\tau, s}^{(l)}, \mathbf{h}_{\tau, q}^{(l)}$  of  $l^{\text{th}}$  GNN layer by Eq. 1;
8:       Obtain molecular representations  $\mathbf{z}_{\tau, s}^{(l)}, \mathbf{z}_{\tau, q}^{(l)}$  of  $l^{\text{th}}$  GNN layer by Eq. 2
9:       Evaluate prediction  $\hat{\mathbf{y}}_{\tau, q}^{(l)}$  of  $l^{\text{th}}$  GNN layer by Eq. 3
10:    end for
11:    Evaluate the final prediction  $\hat{\mathbf{y}}_{\tau, q}$  by Eq. 4
12:  end for
13:  Update  $\theta, \mathbf{w}$  by Eq. 5;
14: end while

```

3.3.2 TRAINING AND INFERENCE

For simplicity, we denote HierMatch as $f_{\theta, \mathbf{w}}$, where θ includes the parameters of all graph-based molecular encoder layers, and $\mathbf{w} = \{\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_o\}$ represents the parameters of the matching and fusion modules.

Training Stage. We employ a standard meta-learning process to train the model on the training set $\mathcal{D}_{\text{train}}$ to enhance its generalization performance. For each training task \mathcal{T}_τ , the model is trained on a small set of labeled samples, known as the support set \mathcal{S}_τ , and then evaluated on the query set \mathcal{Q}_τ . By iteratively training on different tasks within $\mathcal{D}_{\text{train}}$, the model gradually acquires general knowledge across tasks, thereby strengthening its generalization capability. After obtaining the prediction results $\hat{\mathbf{y}}_{\tau, q}$ for the query set using Eq. 4, we optimize the model using the following objective function:

$$\min_{f_{\tau, \theta, \mathbf{w}}} \mathbb{E}_{\mathcal{S}_\tau, \mathcal{Q}_\tau \sim \mathcal{D}_{\text{train}}} \left[\frac{1}{|\mathcal{Q}_\tau|} \sum_{(\mathbf{x}_{\tau, j}, y_{\tau, j}) \in \mathcal{Q}_\tau} \mathcal{L}(y_{\tau, j}, \hat{y}_{\tau, j}) \right], \quad (5)$$

where \mathcal{L} represents the cross-entropy (CE) loss function, and $y_{\tau, j} \in \{0, 1\}$ represents the ground-truth label. The complete meta-training algorithm of HierMatch is detailed in Algorithm 1.

Inference Stage. After training, we evaluate the model using a series of test tasks $\mathcal{T}_{\text{test}}$ on the test set $\mathcal{D}_{\text{test}}$. To capture the specific knowledge of each test task $\mathcal{T}_{\text{test}, \tau}$, we split the support set $\mathcal{S}_{\text{test}, \tau}$ into $\tilde{\mathcal{S}}_\tau$ and $\tilde{\mathcal{Q}}_\tau$. With the parameter θ fixed, we fine-tune the model parameters \mathbf{w} on $\tilde{\mathcal{Q}}_\tau$. The optimization objective function is given as follows:

$$\min_{\mathbf{w}_\tau} \mathbb{E}_{\tilde{\mathcal{S}}_\tau, \tilde{\mathcal{Q}}_\tau \sim \mathcal{S}_{\tau, \text{test}}} \left[\frac{1}{|\tilde{\mathcal{Q}}_\tau|} \sum_{(\mathbf{x}_{\tau, j}, y_{\tau, j}) \in \tilde{\mathcal{Q}}_\tau} \mathcal{L}(y_{\tau, j}, \hat{y}_{\tau, j}) \right], \quad (6)$$

where \mathbf{w}_τ denotes the task-specific parameters for each task. After fine-tuning, the model is evaluated using the support set $\tilde{\mathcal{S}}_{\text{test}, \tau}$ to predict the labels of unknown query molecules. This process allows the model to adapt to the specific features of each test task, enabling accurate predictions for new and unseen molecular structures.

4 EXPERIMENT

In this section, we evaluate the empirical performance of HierMatch, as outlined in Section 3.3. We conduct experiments on the MoleculeNet 4.1 and FS-Mol 4.2 benchmarks to validate our approach.

Additionally, we perform an ablation study of HierMatch in Section 4.2. To demonstrate the generalization of HierMatch, we further test it on seven datasets from the Meta-MolNet benchmark in Section 4.3, covering both single-task and multi-task scenarios. [Lastly, we conduct visualization experiments to demonstrate the importance of hierarchical matching in HierMatch in Section 4.4.](#) For a more detailed set of experimental results, please refer to Appendix D.3. All experiments are run on an NVIDIA RTX A6000 GPU.

4.1 FEW-SHOT MOLECULAR PROPERTY PREDICTION ON MOLECULENET BENCHMARK

Table 1: All methods are compared on the MoleculeNet benchmark with a support set size of 20. The mean test performance measured by AUROC% along with the standard deviations.

Method	Tox21 (12) \uparrow	SIDER (27) \uparrow	MUV (17) \uparrow	ToxCast (617) \uparrow
CHEF (Adler et al., 2020)	61.97 \pm 0.65	57.34 \pm 0.82	53.17 \pm 4.21	56.52 \pm 1.24
MixHop (Abu-El-Haija et al., 2019)	78.14 \pm 0.33	72.01 \pm 0.87	78.04 \pm 3.01	77.19 \pm 0.93
Siamese (Koch et al., 2015)	80.40 \pm 0.35	71.10 \pm 4.32	59.59 \pm 5.13	-
ProtoNet (Snell et al., 2017)	74.98 \pm 0.32	64.54 \pm 0.89	65.88 \pm 4.11	63.70 \pm 1.26
MAML (Finn et al., 2017)	80.21 \pm 0.24	70.43 \pm 0.76	63.90 \pm 2.28	66.79 \pm 0.85
TPN (Liu et al., 2018)	76.05 \pm 0.24	67.84 \pm 0.95	65.22 \pm 5.82	62.74 \pm 1.45
EGNN (Kim et al., 2019)	81.21 \pm 0.16	72.87 \pm 0.73	65.20 \pm 2.08	63.65 \pm 1.57
IterRefLSTM (Altae-Tran et al., 2017)	81.10 \pm 0.17	69.63 \pm 0.31	45.56 \pm 5.12	-
PAR (Wang et al., 2021)	82.06 \pm 0.12	74.68 \pm 0.31	66.48 \pm 2.12	69.72 \pm 1.63
ADKF-IFT (Chen et al., 2023)	82.43 \pm 0.60	67.72 \pm 1.21	98.18 \pm 3.05	72.07 \pm 0.81
MHNfs (Schimunek et al., 2023)	80.23 \pm 0.84	65.89 \pm 1.17	73.81 \pm 2.53	74.91 \pm 0.73
HierMatch (Ours)	82.62 \pm 0.43	68.13 \pm 1.54	79.40 \pm 3.14	77.74 \pm 0.75
Pre-GNN (Hu et al., 2020)	82.14 \pm 0.08	73.96 \pm 0.08	67.14 \pm 1.58	73.68 \pm 0.74
GNN-MAML (Guo et al., 2021)	82.97 \pm 0.10	75.43 \pm 0.21	68.99 \pm 1.84	-
Pre-PAR (Wang et al., 2021)	84.93 \pm 0.11	78.08 \pm 0.16	69.96 \pm 1.37	75.12 \pm 0.84
Pre-ADKF-IFT (Chen et al., 2023)	86.06 \pm 0.35	70.95 \pm 0.60	95.74 \pm 0.37	76.22 \pm 0.13
Pre-HierMatch (Ours)	86.35 \pm 0.13	80.34 \pm 0.45	86.35 \pm 0.76	81.63 \pm 0.73

Benchmark and Baselines. MoleculeNet (Wu et al., 2018) serves as a benchmark for few-shot molecular property prediction, focusing on small molecules with a molecular weight of less than 900 Daltons. This benchmark includes 4 datasets: Tox21, SIDER, MUV, and ToxCast. More details of datasets refer to Appendix A.1. We compare **HierMatch** with two types of baselines: 1) Methods trained from scratch, including **CHEF** (Adler et al., 2020), **MixHop** (Abu-El-Haija et al., 2019), **Siamese** (Koch et al., 2015), **ProtoNet** (Snell et al., 2017), **MAML** (Ren et al., 2018), **TPN** (Liu et al., 2018), **EGNN** (Kim et al., 2019), **IterRefLSTM** (Altae-Tran et al., 2017), **PAR** (Wang et al., 2021), **MHNfs** (Schimunek et al., 2023), and **ADKF-IFT** (Chen et al., 2023); 2) Methods that fine-tune pretrained models, including **Pre-GNN** (Hu et al., 2020), **GNN-MAML** (Guo et al., 2021), **Pre-PAR** (Wang et al., 2021), **Pre-ADKF-IFT** (Chen et al., 2023). **Pre-HierMatch** is our pretrained method, utilizing pretrained parameters provided by the Pre-GNN method (Hu et al., 2020). More details of the experimental setup can be found in Appendix A.2.

Evaluation Procedure. Following the procedural framework of Wang et al. (2021), we adopt AUROC (Area Under the Receiver Operating Characteristic Curve) as the evaluation metric and set the support set to 20 (i.e., 2-way 10-shot). The model is trained using the Adam optimizer (Kingma & Ba, 2014). During testing, results are based on 10 repeated experiments with different random seeds. For the baselines, we replicated the results for CHEF, MixHop and MHNfs, while the results for the other baselines are cited from Chen et al. (2023).

Performance. Table 1 demonstrates that both **HierMatch** and **Pre-HierMatch** outperform current leading methods on the Tox21, SIDER (pre-training stage only), and ToxCast datasets, surpassing the **SOTAs** by an average margin of **2.87%**. [Compared to CHEF, HierMatch demonstrates superior performance, suggesting that graph structures are more effective than fixed fingerprints for hierarchical representation learning in this context.](#) Additionally, our HierMatch outperforms MixHop, highlighting the importance of hierarchical matching for molecular property prediction,

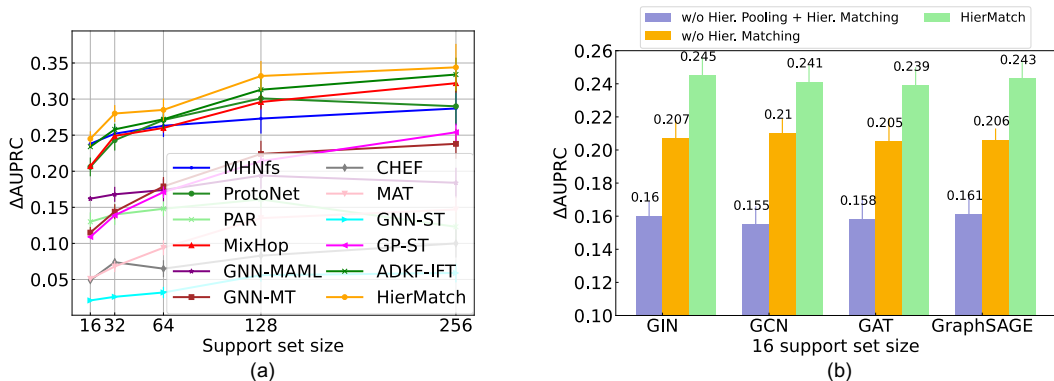


Figure 3: Mean performance with standard errors on the FS-Mol test tasks. (a) Performance of all compared approaches on FS-Mol benchmark. (b) Ablation study of the hierarchical pooling and matching modules in HierMatch across different backbones.

especially in few-shot scenarios. On the MUV dataset, HierMatch ranks second among all base-lines, possibly due to the severe distribution imbalance inherent in the MUV dataset.

4.2 FEW-SHOT MOLECULAR PROPERTY PREDICTION ON FS-MOL BENCHMARK

Benchmark and Baselines. FS-Mol, introduced by Stanley et al. (2021), serves as a benchmark for few-shot molecular property prediction tasks, specifically for macromolecules (i.e., proteins). It comprises 5,120 tasks, divided into a training set (4,938 tasks), a validation set (40 tasks), and a test set (157 tasks), covering a total of 233,786 compounds. Further details are provided in Appendix B.1. For comparison with **HierMatch**, we select four types of baselines: 1) Single-task methods: single-task GP with Tanimoto kernel (**GP-ST**) (Ralaivola et al., 2005), single-task GNN (**GNN-ST**) (Gilmer et al., 2017), and **CHEF** (Adler et al., 2020); 2) Multi-task pre-training method: Multi-task GNN (**GNN-MT**) (Stanley et al., 2021); 3) Self-supervised pre-training method: Molecule Attention Transformer (**MAT**) (Maziarka et al., 2020); and 4) Meta-learning methods: **MixTop** (Abu-El-Haija et al., 2019), **PAR** (Wang et al., 2021), **ProtoNet** (Snell et al., 2017), **GNN-MAML** (Guo et al., 2021), **ADKF-IFT** (Chen et al., 2023) and **MHNfs** (Schimunek et al., 2023). The experimental results of all baselines are reproduced by Chen et al. (2023). Further details can be found in Appendix B.2.

Evaluation Procedure. We adopt the identical experimental configuration as the FS-Mol benchmark (Stanley et al., 2021). For each task, we employ unbalanced sampling to create an uneven distribution of positive and negative samples within the support set. The evaluation metric, Δ AUPRC, provides a precise and sensitive measure of how well the model improves in handling minority classes, which is crucial in unbalanced datasets. Further details of the evaluation can be found in Appendix B.3. During testing, we set five different support set sizes: 16, 32, 64, 128, and 256. For each setting, we perform 10 repeated random splits of the support/query sets for the test tasks under these settings and take the averages as the final results.

Performance. Figure 3 (a) displays the test results of all compared methods. The results indicate that **HierMatch** outperforms all benchmarks across various support set sizes. It achieves substantial performance gains of 4.27%, 8.53%, 4.40%, 6.07%, and 4.26% with support set sizes of 16, 32, 64, 128, and 256, respectively. These findings underscore the effectiveness of HierMatch’s hierarchical pooling and matching mechanism in enhancing the model’s generalization and robustness. Additionally, HierMatch demonstrates strong adaptability, consistently improving performance across different support set sizes.

Ablation Study. 1) To explore the importance of hierarchical pooling and matching mechanisms in learning complex molecular structures, we select GNN-MAML (Guo et al., 2021) as the baseline. HierMatch extends GNN-MAML by incorporating these hierarchical mechanisms. 2) To evaluate

the transferability of HierMatch, we tested it using common GNN frameworks, including GIN (Xu et al., 2019), GCN (Kipf & Welling, 2017), GAT (Velićković et al., 2018), and GraphSAGE (Hamilton et al., 2017). Experimental results, as shown in Figure 3 (b), highlight the significant advantages of the hierarchical pooling and matching mechanism in processing complex molecular structure information and showcase their strong adaptability and transferability across frameworks.

Sub-benchmark Performance. The FS-Mol benchmark (157 test tasks) is divided into 7 subset tasks (Stanley et al., 2021). The results of these subset tasks for HierMatch and the baselines are presented in Appendix D.2. Table D.2 demonstrates the superior performance of HierMatch over SOTA.

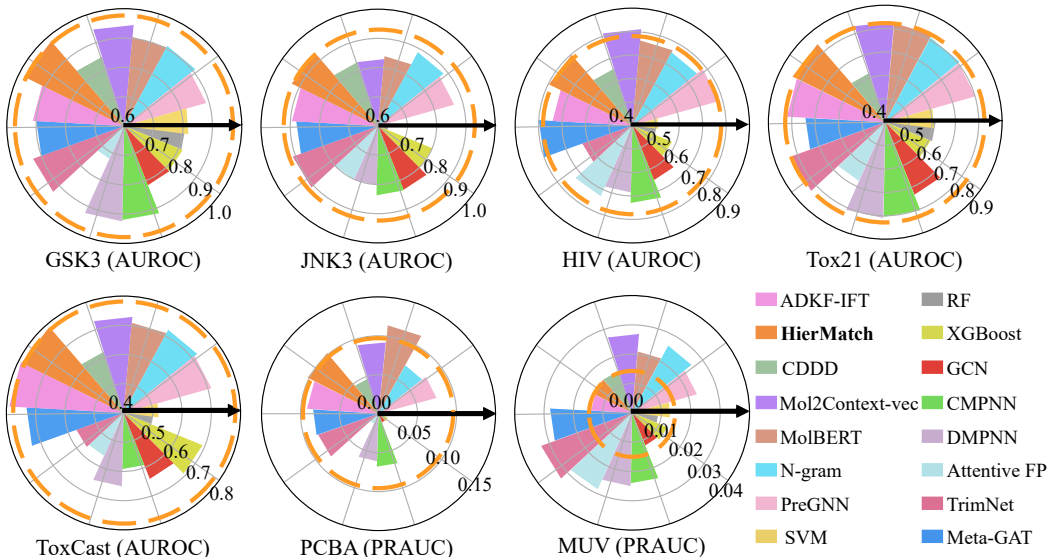


Figure 4: The performance of all compared methods on the seven classification tasks with a support set of size 2 on the Meta-MolNet benchmark. Each colored sector represents a method, with the height of the sector indicating the method’s effectiveness. Starting from the black arrow, the methods are listed in the legend in a counterclockwise direction. **HierMatch** corresponds to the orange sector. The dashed orange circle marks the results of HierMatch. Methods with sectors not crossing this line fail to surpass HierMatch, while those crossing it show superior performance.

4.3 CROSS-DOMAIN DRUG DISCOVERY ON META-MOLNET BENCHMARK

Benchmark and Baselines. Meta-MolNet (Lv et al., 2024) sets a standard for evaluating generalization in computational chemistry by improving data quality and testing rigor. We evaluate our model on classification tasks including GSK3, JNK3, HIV, Tox21, ToxCast, PCBA, and MUV. For comparison, we consider four types of baselines: 1) Classical machine learning methods: support vector machine (SVM) (Bao et al., 2016), extreme gradient boosting algorithm (XGBoost) (Deng et al., 2021), and random forests (RF) (Fabris et al., 2018). 2) Supervised learning methods: GCN (Kipf & Welling, 2016), CMPNN (Song et al., 2020), DMPNN (Yang et al., 2019), Attentive FP (Xiong et al., 2019), and TrimNet (Li et al., 2020). 3) Self-supervised learning methods: CDDD (Winter et al., 2019), Mol2Context-vec (Lv et al., 2021), MolBERT (Fabian et al., 2020), N-gram (Liu et al., 2019), and Pre-GNN (Hu et al., 2020). 4) Meta-learning method: ADKF-IFT (Chen et al., 2023) and Meta-GAT (Lv et al., 2024). All baseline results are reproduced according to Lv et al. (2024). Due to the sub-task settings of Meta-MolNet, prototype-based methods are no longer applicable. Further details can be found in Appendix C.1.

Evaluation Procedure. To evaluate the generalization ability of HierMatch, we follow a higher ratio of molecules/scaffolds as Lv et al. (2024). For classification tasks, we use AUROC and PRAUC as evaluation metrics. Specifically, AUROC is used to measure the performance of binary classification tasks (GSK3, JNK3, HIV, Tox21, and ToxCast), while PRAUC is more suitable for tasks with

severely skewed distributions (PCBA, MUV). All experimental results are based on the average of three independent runs with different random seeds, with the support set size set to 2. Further details on evaluation metrics can be found in Appendix C.3.

Performance. Figure 4 shows the comparative results of different methods on the seven classification datasets in Meta-MolNet. The experiments indicate that HierMatch performs excellently on the GSK3, JNK3, Tox21, and ToxCast datasets, while showing less well on the HIV and PCBA datasets. Our method faces significant challenges on the MUV dataset, likely due to distributional biases. Overall, HierMatch exhibits excellent generalization capabilities across most datasets for new molecular scaffolds but performs poorly in specific cases, such as the MUV dataset.

4.4 VISUALIZATION

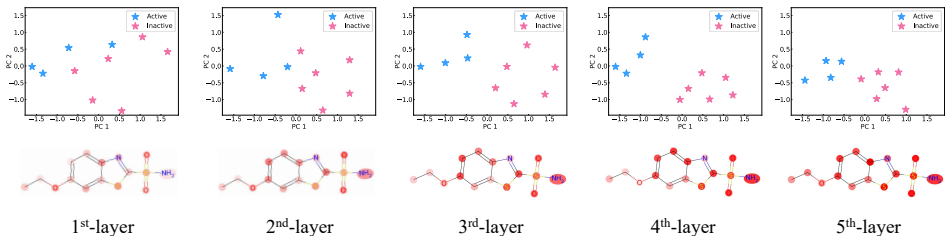


Figure 5: Layer-wise visualization for NR-AhR toxicity prediction. The first row presents PCA projections of 10 molecules, distinguishing between active (blue) and inactive (pink) molecules. The second row displays the internal visualization of a selected molecule across layers, with color intensity indicating the model’s attention shifts as the layers deepen.

To validate the importance of hierarchical representations, we visualize 10 molecules for the NR-AhR toxicity prediction task, as shown in Figure 5. In the second row, we select one molecule, with the SMILES CCOc1ccc2nc(S(N)(=O)=O)sc2c1, to demonstrate how each GNN layer captures distinct structural levels, from atoms and substructures to the entire molecule. Additionally, PCA projections of the 10 molecules were performed to examine the distribution of active and inactive compounds. This analysis enhances our understanding of the model’s ability to distinguish molecular structures across layers, offering insights into both its interpretability and the role of hierarchical feature extraction in toxicity prediction. Further details can be found in the Appendix E.

5 CONCLUSION

We propose Hierarchical Matching Networks (HierMatch) to address the limitations of existing few-shot learning methods in drug discovery. HierMatch involves hierarchical pooling and matching techniques. Specifically, hierarchical pooling aggregates node features at each level, preserving crucial details that single-scale methods might overlook. Then, hierarchical matching performs pairwise comparisons at these different structural levels, allowing more accurate assessments of molecular similarities. Experimental results show that HierMatch improves AUROC and Δ AUPRC by 2.87% and 6.52% respectively on the MoleculeNet and FS-Mol benchmarks and demonstrates excellent generalization on the Meta-MolNet benchmark. Future work will focus on improving the fusion mechanism of HierMatch by adopting advanced techniques such as attention fusion or multi-scale feature aggregation to better capture the complex relationships between structural levels.

6 ETHICS STATEMENT

This paper addresses few-shot molecular representation learning without introducing new datasets or requiring human annotation. As such, it presents no additional ethical concerns beyond those commonly associated with research in this field.

REFERENCES

- Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *international conference on machine learning*, pp. 21–29. PMLR, 2019.
- Temitope Isaac Adelusi, Abdul-Quddus Kehinde Oyedele, Ibrahim Damilare Boyenle, Abdeen Tunde Ogunlana, Rofiat Oluwabusola Adeyemi, Chiamaka Divine Ukachi, Mukhtar Oluwaseun Idris, Olamide Tosin Olaoba, Ibrahim Olaide Adedotun, Oladipo Elijah Kolawole, et al. Molecular modeling in drug discovery. *Informatics in Medicine Unlocked*, 29: 100880, 2022.
- Thomas Adler, Johannes Brandstetter, Michael Widrich, Andreas Mayr, David Kreil, Michael K Kopp, Günter Klambauer, and Sepp Hochreiter. Cross-domain few-shot learning by representation fusion. 2020.
- Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.
- Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021.
- Yu Bao, Morihiro Hayashida, and Tatsuya Akutsu. Lbsizecleav: improved support vector machine (svm)-based prediction of dicer cleavage sites using loop/bulge length. *BMC bioinformatics*, 17: 1–11, 2016.
- Andreas Bender, Hamse Y Mussa, Robert C Glen, and Stephan Reiling. Similarity searching of chemical databases using atom environment descriptors (molprint 2d): evaluation of performance. *Journal of chemical information and computer sciences*, 44(5):1708–1718, 2004.
- Wenlin Chen, Austin Tripp, and José Miguel Hernández-Lobato. Meta-learning adaptive deep kernel gaussian processes for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=KXRSh0sdVTP>.
- Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.
- Daiguo Deng, Xiaowei Chen, Ruochi Zhang, Zengrong Lei, Xiaojian Wang, and Fengfeng Zhou. Xgraphboost: extracting graph neural network-based features for a better prediction of molecular properties. *Journal of chemical information and modeling*, 61(6):2697–2705, 2021.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Jurgen Drews. Drug discovery: a historical perspective. *science*, 287(5460):1960–1964, 2000.
- David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.
- Benedek Fabian, Thomas Edlich, Hélène Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.
- Fabio Fabris, Aoife Doherty, Daniel Palmer, Joao Pedro de Magalhaes, and Alex A Freitas. A new approach for interpreting random forest models and its application to the biology of ageing. *Bioinformatics*, 34(14):2449–2456, 2018.

- Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428, 2019.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning, pp. 1126–1135. PMLR, 2017.
- Erik Gawehn, Jan A Hiss, and Gisbert Schneider. Deep learning in drug discovery. Molecular informatics, 35(1):3–14, 2016.
- 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.
- Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. IDrugs, 9(3):199, 2006.
- Daniele Grattarola, Daniele Zambon, Filippo Maria Bianchi, and Cesare Alippi. Understanding pooling in graph neural networks. IEEE transactions on neural networks and learning systems, 2022.
- Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. Few-shot graph learning for molecular property prediction. In The Web Conference 2021-Proceedings of the World Wide Web Conference, WWW 2021, pp. 2559–2567. ACM, 2021.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. Advances in neural information processing systems, 30, 2017.
- Qian-Nan Hu, Hui Zhu, Xiaobing Li, Manman Zhang, Zhe Deng, Xiaoyan Yang, and Zixin Deng. Assignment of ec numbers to enzymatic reactions with reaction difference fingerprints. PloS one, 7(12):e52901, 2012.
- W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. Strategies for pre-training graph neural networks. In International Conference on Learning Representations (ICLR), 2020.
- Donggyun Kim, Jinwoo Kim, Seongwoong Cho, Chong Luo, and Seunghoon Hong. Universal few-shot learning of dense prediction tasks with visual token matching. In The Eleventh International Conference on Learning Representations, 2023. URL <https://openreview.net/forum?id=88nT0j5jAn>.
- Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11–20, 2019.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations, 2016.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations, 2017. URL <https://openreview.net/forum?id=SJU4ayYgl>.
- Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2. Lille, 2015.
- Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. Nucleic acids research, 44(D1):D1075–D1079, 2016.
- Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In International conference on machine learning, pp. 3734–3743. PMLR, 2019.

- Huaiyu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, and Bao-Gang Hu. Lgm-net: Learning to generate matching networks for few-shot learning. In International conference on machine learning, pp. 3825–3834. PMLR, 2019.
- Pengyong Li, Yuquan Li, Chang-Yu Hsieh, Shengyu Zhang, Xianggen Liu, Huanxiang Liu, Sen Song, and Xiaojun Yao. TrimNet: learning molecular representation from triplet messages for biomedicine. Briefings in Bioinformatics, 22(4):bbaa266, 11 2020. ISSN 1477-4054. doi: 10.1093/bib/bbaa266. URL <https://doi.org/10.1093/bib/bbaa266>.
- Xiaoxu Li, Xiaochen Yang, Zhanyu Ma, and Jing-Hao Xue. Deep metric learning for few-shot image classification: A review of recent developments. Pattern Recognition, 138:109381, 2023.
- Peixun Liu and Wei Long. Current mathematical methods used in qsar/qspr studies. International Journal of Molecular Sciences, 10(5):1978–1998, 2009.
- Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. Advances in neural information processing systems, 32, 2019.
- Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002, 2018.
- Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. arXiv preprint arXiv:2406.12950, 2024.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
- Qiuji Lv, Guanxing Chen, Lu Zhao, Weihe Zhong, and Calvin Yu-Chian Chen. Mol2Context-vec: learning molecular representation from context awareness for drug discovery. Briefings in Bioinformatics, 22(6):bbab317, 08 2021. doi: 10.1093/bib/bbab317. URL <https://doi.org/10.1093/bib/bbab317>.
- Qiuji Lv, Guanxing Chen, Ziduo Yang, Weihe Zhong, and Calvin Yu-Chian Chen. Meta-molnet: A cross-domain benchmark for few examples drug discovery. IEEE Transactions on Neural Networks and Learning Systems, 2024.
- Stephani Joy Y Macalino, Vijayakumar Gosu, Sunhye Hong, and Sun Choi. Role of computer-aided drug design in modern drug discovery. Archives of pharmacal research, 38:1686–1701, 2015.
- Sijie Mai, Haifeng Hu, and Jia Xu. Attentive matching network for few-shot learning. Computer Vision and Image Understanding, 187:102781, 2019.
- Kit-Kay Mak, Yi-Hang Wong, and Mallikarjuna Rao Pichika. Artificial intelligence in drug discovery and development. Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays, pp. 1–38, 2023.
- Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. arXiv preprint arXiv:2002.08264, 2020.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.
- Liva Ralaivola, Sanjay J Swamidass, Hiroto Saigo, and Pierre Baldi. Graph kernels for chemical informatics. Neural networks, 18(8):1093–1110, 2005.
- Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. arXiv preprint arXiv:2008.02217, 2020.

- Ekagra Ranjan, Soumya Sanyal, and Partha Talukdar. Asap: Adaptive structure aware pooling for learning hierarchical graph representations. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pp. 5470–5477, 2020.
- Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In International Conference on Learning Representations, 2018.
- Jean-Paul Renaud, Chun-wa Chung, U Helena Danielson, Ursula Egner, Michael Hennig, Roderick E Hubbard, and Herbert Nar. Biophysics in drug discovery: impact, challenges and opportunities. Nature reviews Drug discovery, 15(10):679–698, 2016.
- Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, John F Wambaugh, et al. Toxcast chemical landscape: paving the road to 21st century toxicology. Chemical research in toxicology, 29(8):1225–1251, 2016.
- Ann M Richard, Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J Collins, Inthirany Thillainadarajah, Christopher M Grulke, Antony J Williams, Ryan R Lougee, Richard S Judson, et al. The tox21 10k compound library: collaborative chemistry advancing toxicology. Chemical Research in Toxicology, 34(2):189–216, 2020.
- David Rogers and Mathew Hahn. Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5):742–754, 2010.
- Sebastian G Rohrer and Knut Baumann. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. Journal of chemical information and modeling, 49(2):169–184, 2009.
- Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. In The Eleventh International Conference on Learning Representations, 2023. URL <https://openreview.net/forum?id=XrMWUUEvr>.
- Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe. Computational methods in drug discovery. Pharmacological reviews, 66(1):334–395, 2014.
- Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. Advances in neural information processing systems, 30, 2017.
- Ying Song, Shuangjia Zheng, Zhangming Niu, Zhang-Hua Fu, Yutong Lu, and Yuedong Yang. Communicative representation learning on attributed molecular graphs. In IJCAI, volume 2020, pp. 2831–2838, 2020.
- Megan Stanley, John F Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-mol: A few-shot learning dataset of molecules. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. URL <https://openreview.net/forum?id=701FtuyLlAd>.
- Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1199–1208, 2018.
- Alexander Tropsha, Olexandr Isayev, Alexandre Varnek, Gisbert Schneider, and Artem Cherkasov. Integrating qsar modelling and deep learning in drug discovery: the emergence of deep qsar. Nature Reviews Drug Discovery, pp. 1–15, 2023.
- Thomas Unterthiner, Andreas Mayr, Günter Klambauer, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, and Sepp Hochreiter. Deep learning as an opportunity in virtual screening. In Proceedings of the deep learning workshop at NIPS, volume 27, pp. 1–9, 2014.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In International Conference on Learning Representations, 2018.
- Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. Advances in neural information processing systems, 29, 2016.
- Ning Wang, Minnan Luo, Kaize Ding, Lingling Zhang, Jundong Li, and Qinghua Zheng. Graph few-shot learning with attribute matching. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1545–1554, 2020a.
- Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3):1–34, 2020b.
- Yaqing Wang, ABULIKEMU ABUDUWEILI, quanming yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), Advances in Neural Information Processing Systems, 2021. URL <https://openreview.net/forum?id=vGjTOxss-Dl>.
- Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chemical science, 10(6):1692–1701, 2019.
- 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.
- Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of medicinal chemistry, 63(16):8749–8760, 2019.
- Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In International conference on machine learning, pp. 5453–5462. PMLR, 2018.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In International Conference on Learning Representations, 2019. URL <https://openreview.net/forum?id=ryGs6iA5Km>.
- 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.
- 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.
- Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. Communications Chemistry, 6(1):34, 2023.
- 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, 2021a. URL <https://openreview.net/forum?id=to9OkIFYyC>.
- Zhen Zhang, Jiajun Bu, Martin Ester, Jianfeng Zhang, Chengwei Yao, Zhi Yu, and Can Wang. Hierarchical graph pooling with structure learning. arXiv preprint arXiv:1911.05954, 2019.

- Ziqiao Zhang, Jihong Guan, and Shuigeng Zhou. Fragat: a fragment-oriented multi-scale graph attention model for molecular property prediction. *Bioinformatics*, 37(18):2981–2987, 2021b.
- Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*, 36:5850–5887, 2023.
- Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *European Conference on Computer Vision*, pp. 297–313. Springer, 2022.
- Yangqing Zhong, Yuling Su, and Hong Zhao. Self-similarity feature based few-shot learning via hierarchical relation network. *International Journal of Machine Learning and Cybernetics*, 14(12):4237–4249, 2023.
- Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.
- Xiang Zhuang, Qiang Zhang, Bin Wu, Keyan Ding, Yin Fang, and Huajun Chen. Graph sampling-based meta-learning for molecular property prediction. *arXiv preprint arXiv:2306.16780*, 2023.

A DETAILS OF MOLECULENET BENCHMARK

In this section, we introduce the details of datasets that are included in the MoleculeNet benchmark in Section A.1. In addition, we show the details of the experimental setup A.2.

A.1 DETAILS OF DATASETS

Table 2: Summary of datasets included in MoleculeNet.

Dataset	Tox21	SIDER	MUV	ToxCast
Compounds	8,014	1,427	93,127	8,615
Tasks	12	27	17	617
Meta-Training Tasks	9	21	12	450
Meta-Testing Tasks	3	6	5	167

In the MoleculeNet benchmark, we perform experiments on 4 datasets in Tabel 2, which include Tox21 (Richard et al., 2020), SIDER (Kuhn et al., 2016), MUV (Rohrer & Baumann, 2009), and ToxCast (Richard et al., 2016). Widely utilized in the assessment of compound toxicity for drug development and environmental risk evaluation, the Tox21 dataset, as described in Richard et al. (2020), contains 8,014 compounds categorized into 12 tasks. By analyzing this dataset, researchers can identify environmental pollutants and potential drug candidates, offering crucial insights into their impact on human health. The SIDER dataset, introduced in Kuhn et al. (2016), serves as a crucial database of drug side effects, encompassing extensive information on medications and their associated adverse responses. This dataset encompasses 1427 compounds distributed among 27 categories. Utilizing the SIDER dataset provides researchers with valuable insights into drug safety profiles and potential side effects. The MUV dataset (Rohrer & Baumann, 2009), which includes 93,127 compounds distributed among 17 tasks showcasing a range of biological activities, is widely acknowledged as a key standard for evaluating the multifaceted functions of drug compounds. A fundamental resource in toxicology research, the ToxCast dataset (Richard et al., 2016) is a critical high-throughput screening database used to evaluate the potential health hazards posed by various compounds. With a compilation of 8,615 compounds and 617 tasks, this dataset significantly contributes to the field of toxicology.

A.2 DETAILS OF EXPERIMENTAL SETUP

In HieMatch (and Pre-HirMatch), GIN used in Equation 1 and Equation 2 consists of 5 layers with hidden size 300. In addition, attention mechanism used in Equation 3 consist of 1 layer with 1 head. We implement HierMatch in PyTorch (Paszke et al., 2019) and Pytorch Geometric library (Fey & Lenssen, 2019). We train the model for a maximum number of 5000 epoches. We employ the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 for meta-learning, while using a higher learning rate of 0.05 for fine-tuning the matching module and fusion module within each task. The dropout rate is maintained at 0.1 for all components, except for the graph-based molecular encoder. We summarize the hyperparameters used by HierMatch in Table 3

Table 3: Hyperparameters used by HierMatch

Hyperparameter	Explored values	Selected
learning rate for meta-learning	0.001	0.001
learning rate for fine-tuning	0.01~0.5	0.05
number of update steps for fine-tuning	1~5	5
number of layer of GNN in (1) and (2)	5	5
number of layer of matching module in (3)	1	1
number of head of matching module in (3)	1	1
dropout	0.0~0.5	0.1
hidden dimension for GNN in (1) and (2)	300	300

B DETAILS OF FS-MOL BENCHMARK

In this section, we first introduce the details of FS-Mol benchmark (Stanley et al., 2021) in Section B.1. The subsequent discussion delves into the details of the compared baselines on FS-Mol benchmark in Section B.1. In addition, further details regarding the evaluation metric ΔAUCPR is presented in Section B.3. Finally, the details of experimental setup on FS-Mol benchmark is presented in Section B.4.

B.1 DETAILS OF BENCHMARKS

The Few-Shot Learning Dataset of Molecules (**FS-Mol**) (Stanley et al., 2021) is designed for machine learning applications in the Quantitative Structure-Activity Relationships (QSAR) field (Tropsha et al., 2023), specifically focusing on few-shot learning scenarios. It comprises a total of 5120 distinct assays, encompassing 233,786 unique compounds. The dataset is partitioned into three subsets: D_{train} for training, D_{test} for testing, and D_{valid} for validation purposes. D_{test} contains 157 tasks, D_{train} includes 4938 tasks, and D_{valid} is composed of 40 tasks. Notably, each task in the dataset contains an average of 94 compounds, a notably lower figure compared to other similar datasets. This characteristic reflects the high specificity of the protein targets and the corresponding assays, posing a significant challenge in the QSAR domain.

B.2 DETAILS OF BASELINES.

In the comparative analysis of the FS-Mol benchmark (Stanley et al., 2021), four types of baselines have been chosen: Single-task methods, Multi-task pre-training methods, Self-supervised pre-training methods, and Meta-learning methods.

Single-task Methods. The single-task methods are single-task GP with Tanimoto kernel (**GP-ST**) (Ralaivola et al., 2005), single-task GNN (**GNN-ST**) (Gilmer et al., 2017), and **MHNfs** (Schimunek et al., 2023) for context-enriched information.

GP-ST, as delineated in the study by (Ralaivola et al., 2005), encompassing the random walk kernel, shortest-path kernel, and subtree kernel, are employed to evaluate the resemblance between graphs of chemical compounds. Gilmer et al. (2017) introduces **GNN-ST**, particularly focusing on MPNNs for proficient learning from graph-based representations of molecules in quantum chemistry.

Multi-task Pre-training Method. Multi-task GNN (**GNN-MT**) (Stanley et al., 2021) employs a 10-layer pre-trained GNN with 128 hidden dimensions and "principal neighborhood message aggregation." Task-specific readout functions and an MLP with a 512-dimensional hidden layer produce activity label predictions. The model is fine-tuned on all tasks in \mathcal{D}_{train} using multi-task learning.

Self-supervised Pre-training Method. The Molecule Attention Transformer (**MAT**) (Maziarka et al., 2020) modifies the Transformer architecture (Vaswani et al., 2017) by incorporating insights on inter-atomic distances and the molecular graph structure into the self-attention mechanism.

Meta-learning Methods. Property-Aware Relation Networks (**PAR**) (Wang et al., 2021), Prototypical Networks (**ProtoNet**) (Snell et al., 2017), **GNN-MAML** (Guo et al., 2021), and **ADKF-IFT** (Chen et al., 2023) are four typical meta-learning methods. Specifically, **PAR** (Wang et al., 2021), introduces a property-aware embedding function that transforms generic molecular embeddings into a substructure-aware representation which relevant to the target property, and designs an adaptive relation graph learning module to jointly estimate the molecular relation graph and refine the molecular embeddings with respect to the target property. Schimunek et al. (2023) proposes **MHNfs** approach, utilizing a Modern Hopfield Network (MHN) (Ramsauer et al., 2020) to link molecules with an extensive array of reference molecules, thereby enhancing the covariance structure of the data and mitigating spurious correlations of molecules. **ProtoNet** (Snell et al., 2017), a simple approach to few-shot classification, learns an embedding where each class is represented by a prototype, computed as the mean of the embedded support examples for that class. Classification is then done by computing distances from the query example to each class prototype. **GNN-MAML** (Guo et al., 2021) uses graph neural networks to learn molecular representations, and employs a meta-learning framework for model optimization. It also incorporates molecular structure, self-supervised modules, and self-attentive task weights to exploit unlabeled data and address task heterogeneity. **ADKF-IFT** (Chen et al., 2023) combines the representational power of deep learning with the probabilistic modeling capabilities of gaussian processes, enabling efficient and uncertainty-aware molecular property prediction through meta-learning.

B.3 EVALUATION METRICS OF FS-MOL BENCHMARK

The Δ AUCPR (Area Under the Curve for Precision-Recall) serves as a pivotal statistical measure utilized for assessing enhancements in the efficacy of classification models when confronted with imbalanced datasets due to targeted modifications, like algorithmic adjustments or alterations in data processing methodologies. By contrasting the precision-recall curve’s area prior to and post adjustments, this metric adeptly elucidates the extent of enhancement in the capacity of model to identify minority classes, thereby supplying a quantitative foundation for optimizing the model and facilitating decision-making support.

In line with the research conducted by Stanley et al. (2021), we employ the Δ AUCPR as an evaluation metric for comparing all baseline models. The specific calculation formula is detailed below:

$$\Delta \text{AUCPR}(f_{\theta, \mathbf{w}}) = \text{AUCPR}(f_{\theta, \mathbf{w}}) - \frac{N_{\tau}^q(1)}{N_{\tau}^q}, \quad (7)$$

where the $N_{\tau}^q(1)$ represents the number of active molecules in query set \mathcal{Q}_{τ} .

B.4 DETAILS OF EXPERIMENTAL SETUP

In HierMatch, the hyperparameters used by HierMatch are reported in Table 3. What is more, on FS-Mol benchmark (Stanley et al., 2021), we set the batch task 21 and weight decay $5e-5$. And we train the model for 10,000 epoches.

C DETAILS OF META-MOLNET BENCHMARK

In this section, we first introduce the details of Meta-MolNet benchmark (Lv et al., 2024) in Section C.1. In addition, we provide the details of the baselines in Section C.2. Finally, the details of evaluation metric is provided in Section C.3.

C.1 DETAILS OF BENCHMARKS

Meta-MolNet is an innovative benchmarking platform designed to improve molecular machine learning models by integrating diverse datasets through multitask and transfer learning, spanning applications from drug discovery to materials science. In this paper, we use 7 classification tasks on Meta-MolNet benchmark to evaluate our HierMatch, which include GSK3, JNK3, HIV, Tox21, ToxCast, PCBA and MUV. The GSK3 dataset focuses on predicting the activity of compounds against the GSK3 enzyme, which is associated with diseases like diabetes and Alzheimer’s. The JNK3 dataset assesses the inhibitory activity of compounds against JNK3, a kinase implicated in neurodegenerative diseases. The HIV dataset contains data for predicting the ability of compounds to inhibit HIV replication. Tox21 evaluates the toxicity of compounds across multiple biological pathways, while ToxCast predicts the toxic effects of environmental chemicals. The PCBA dataset measures compound activity across various bioassays from the PubChem database. Lastly, the MUV dataset provides a rigorous and unbiased benchmark for validating virtual screening methods. Together, these tasks offer a comprehensive evaluation framework for molecular machine learning models. The detailed description of datasets in Table 4.

Table 4: Detailed Description of the benchmark datasets

Task type	Datasets	Category	Data type	Tasks	No. of Molecules	No. of Scaffolds	Molecules/ Scaffolds ratio	Metrics	Threshold
Single Task Classification	GSK3	Biophysics	SMILES	1	3,197	38	84.13	ROC-AUC	30
	JNK3	Biophysics	SMILES	1	4,873	62	78.60	ROC-AUC	30
	HIV	Biophysics	SMILES	1	6,386	68	93.91	ROC-AUC	30
Multi Task Classification	Tox21	Physiology	SMILES	12	2,119	12	176.58	ROC-AUC	30
	ToxCast	Physiology	SMILES	617	2,372	14	169.43	ROC-AUC	30
	PCBA	Biophysics	SMILES	128	21,835	34	642.21	PRC-AUC	200
	MUV	Biophysics	SMILES	17	11,671	152	76.78	PRC-AUC	30

C.2 DETAILS OF BASELINES

Four types of baselines—classical machine learning models, graph-based models, message passing neural networks, and self-supervised pre-training models—are chosen for comparative analysis on the Meta-MolNet benchmark (Lv et al., 2024).

Classical Machine Learning Methods. Support Vector Machines (**SVM**) (Bao et al., 2016), extreme gradient boosting algorithms (**XGBoost**) (Deng et al., 2021), and Random Forests (**RF**) (Fabris et al., 2018) are among the classical machine learning methods that utilize descriptors and/or fingerprints commonly found in traditional QSPR/QSAR models (Cherkasov et al., 2014). Notably, the Extended Connectivity Fingerprints (ECFPs) (Rogers & Hahn, 2010; Glen et al., 2006) and Molecular ACCess System (MACCS) keys (Bender et al., 2004; Unterthiner et al., 2014) are widely used as fingerprints in such models. **SVM** (Bao et al., 2016) is a robust machine learning algorithm designed to identify the optimal solution for classification tasks by determining the maximum margin hyperplane within a high-dimensional space. **XGBoost** (Deng et al., 2021) is a proficient machine learning technique that utilizes distributed gradient boosting to provide rapid, adaptable, and user-friendly solutions. Information about **RF** (Fabris et al., 2018) can be found in Appendix B.1.

Supervised Learning Methods. Graph Convolutional Networks (**GCN**) (Duvenaud et al., 2015), Directed Message Passing Neural Networks (**DMPNN**) (Yang et al., 2019), Communicative Message Passing Neural Networks (**CMPNN**) (Song et al., 2020), **Attentive FP** (Xiong et al., 2019), and Triplet Message Networks (**TrimNet**) (Li et al., 2020) are among the supervised learning methods. Specifically, **GCN**³ (Duvenaud et al., 2015) employs convolution operations based on the eigen decomposition of the Laplacian matrix, which allows them to aggregate information from neighboring nodes and derive node embedding representations. **DMPNN**⁴ (Yang et al., 2019) use Laplacian eigen decomposition for convolution operations, aggregating information from neighboring nodes to derive node embeddings. **CMPNN**⁵ (Song et al., 2020) enhance modeling of molecular properties

³<https://github.com/tkipf/gcn.git>

⁴<https://github.com/chemprop/chemprop.git>

⁵<https://github.com/SY575/CMPNN.git>

by using a node-edge interaction module to effectively integrate atom and bond features. **Attentive FP**⁶ (Xiong et al., 2019) employs atom and bond attributes to create feature vectors, preserving spatial information and capturing both local and nonlocal effects with a graph attention mechanism. **TrimNet**⁷ (Li et al., 2020) utilizes a triplet message mechanism to extract edge information from atom-bond-atom interactions, achieving state-of-the-art performance.

Self-supervised Learning Methods. **CDDD** (Winter et al., 2019), **Mol2Context-vec** (Lv et al., 2021), **MolBERT**, **N-gram**, and **Pre-GNN** (Hu et al., 2020) are self-supervised methods that pre-train on large molecular datasets to extract meaningful descriptors. These data-driven approaches produce generalizable features, avoiding fixed extraction rules and reducing overfitting. Specifically, **CDDD**⁸ (Winter et al., 2019) learns features from a large chemical structure corpus by translating between different molecular representations, compressing shared information into a low-dimensional vector. **Mol2Context-vec**⁹ (Lv et al., 2021) uses a Bi-LSTM to create dynamic representations of molecular substructures, capturing intramolecular hydrogen bonds and other non-covalent interactions. **MolBERT**¹⁰ (Fabian et al., 2020) is a Transformer-based model that uses BERT (Devlin et al., 2018) to learn high-quality molecular representations for drug discovery. **N-gram**¹¹ (Liu et al., 2019) captures co-occurrence patterns of local substructures by extracting n-grams from the graph and creating a histogram to represent their frequencies, forming the graph-level representation. **Pre-GNN**¹² (Hu et al., 2020) pre-trains graph neural networks by learning representations at both node and graph levels, capturing local and global structural information in molecular graphs.

Meta-learning Methods. **Meta-GAT** (Lv et al., 2024) and **ADKF-IFT** (Chen et al., 2023) are two typical meta-learning methods. Specifically, **Meta-GAT**¹³ (Lv et al., 2024) is a graph attention network that uses cross-domain meta-learning to predict molecular properties with few examples. By extracting meta-knowledge from similar molecules across domains, it reduces sample complexity and quickly adapts to new scaffold molecules with minimal data. **ADKF-IFT**¹⁴ (Chen et al., 2023) can be seen in Appendix B.1.

C.3 EVALUATION METRICS OF META-MOLNET

In this paper, we use benchmark datasets with a higher ratio of molecules to scaffolds, presenting a significantly more challenging scenario compared to random cross-validation and datasets with a lower ratio (Lv et al., 2024), for evaluating generalization ability. For classification tasks, we use Area Under the Receiver Operating Characteristic Curve (**AUROC**) and Area Under the Precision-Recall Curve (**PRAUC**) as evaluation metrics. Specifically, **AUROC** measures the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) across different classification thresholds. **AUROC** ranges from 0 to 1, where 0.5 represents a random classifier and 1 represents a perfect classifier. A higher **AUROC** value indicates better classification performance, making it well-suited for evaluating binary classification tasks such as GSK3, JNK3, HIV, Tox21, and ToxCast. Meanwhile, **PRAUC** considers the trade-off between precision (positive predictive value) and recall (sensitivity). Like **AUROC**, **PRAUC** ranges from 0 to 1, with higher values indicating better performance. **PRAUC** is particularly useful for evaluating models on imbalanced datasets, making it more suitable for tasks such as PCBA and MUV, which have severely skewed distributions.

C.4 DETAILS OF EXPERIMENTAL SETUP

On the Meta-MolNet benchmark, we set the query set size to 8 and the support set size to 2. We employ the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 0.001 for meta-

⁶<https://github.com/OpenDrugAI/AttentiveFP.git>

⁷<https://github.com/yvquanli/trimnet.git>

⁸<https://github.com/jrwnter/cddd.git>

⁹<https://github.com/lol88/Mol2Context-vec.git>

¹⁰<https://github.com/BenevolentAI/MolBERT.git>

¹¹https://github.com/chaol224/n_gram_graph.git

¹²<https://github.com/snap-stanford/pretrain-gnns.git>

¹³<https://github.com/lol88/Meta-MolNet.git>

¹⁴<https://github.com/Wenlin-Chen/ADKF-IFT.git>

learning and an inner learning rate of 0.001 for fine-tuning the task-specific modules within each task. A weight decay of $5e-4$ is applied. The model is trained for 100 epochs to ensure robust performance.

D FURTHER EXPERIMENTS RESULTS ON FS-MOL

D.1 OVERALL PERFORMANCE

Figure 6 (a)~(e) show the performance of different methods in classifying 157 FS-Mol (Stanley et al., 2021) test tasks across various support set sizes via box plots. The box plots show the distribution of classification accuracies for each method, providing insight into their overall performance and effectiveness in handling varying support set sizes. Our HieMatch demonstrates superior performance compared to the state-of-the-art (SOTA) method across all metrics.

D.2 SUB-BENCHMARK PERFORMANCE

FS-Mol (Stanley et al., 2021) divides tasks into 7 sub-benchmarks using Enzyme Commission (EC) numbers (Hu et al., 2012), allowing for assessment across the entire benchmark. In classification tasks with a support set size of 16, Table D.2 illustrates the performance of the top methods across all sub-benchmarks. The results highlight that, while excelling in overall performance, HieMatch emerges as the top performer in half of the sub-benchmarks for classification tasks.

Table 5: The classification performance for the 16 support set size.

FS-Mol sub-benchmark (EC category)			Method				
Class	Description	#tasks	RF	GP-ST	GNN-MAML	ADKF-IFT	HierMatch
1	oxidoreductases	7	0.081 ± 0.032	0.013 ± 0.019	0.046 ± 0.023	0.103 ± 0.0036	0.231 ± 0.075
2	kinases	125	0.082 ± 0.006	0.013 ± 0.004	0.178 ± 0.009	0.247 ± 0.010	0.256 ± 0.012
3	hydrolases	20	0.158 ± 0.026	0.062 ± 0.019	0.106 ± 0.024	0.213 ± 0.029	0.201 ± 0.028
4	lysases	2	0.218 ± 0.172	0.161 ± 0.112	0.218 ± 0.147	0.223 ± 0.160	0.211 ± 0.061
5	isomerases	1	0.119 ± 0.029	-0.014 ± 0.015	0.006 ± 0.021	0.121 ± 0.049	0.087 ± 0.025
6	ligases	1	0.027 ± 0.069	-0.011 ± 0.003	0.001 ± 0.017	0.103 ± 0.066	0.359 ± 0.011
7	translocases	1	0.102 ± 0.053	0.067 ± 0.050	0.001 ± 0.021	0.082 ± 0.049	-0.009 ± 0.011
all enzymes		157	0.093 ± 0.007	0.021 ± 0.005	0.162 ± 0.009	0.230 ± 0.009	0.245 ± 0.011

D.3 META-TESTING COSTS

In this section, we compare the inference time of our HierMatch with meta-learning approaches. Figure 7 illustrates that HierMatch takes slightly more time compared to ProtoNet (Snell et al., 2017) and GNN-MAML (Guo et al., 2021). Additionally, ADKF-IFT (Chen et al., 2023) exhibits the longest reference time. However, it is important to note that HierMatch still maintains a relatively fast inference time, making it a viable option for meta-learning tasks.

E VISUALIZATION EXPERIMENTS

F DISCUSSION, LIMITATION AND FUTURE WORK

Limitation: Simple Fusion Design. The fusion mechanism in the proposed HierMatch model is relatively simplistic, which might limit its ability to effectively integrate information from different hierarchical levels. This simple design could lead to suboptimal performance as the model may not fully capture the complex interactions and dependencies across multiple scales of molecular structures. A more sophisticated fusion technique, such as attention-based fusion or multi-scale feature aggregation, could potentially enhance the model’s capability to combine features from different layers more effectively. By leveraging advanced fusion strategies, the model can better exploit the rich hierarchical information inherent in molecular structures, leading to improved prediction accuracy and generalization. Implementing these advanced fusion methods may involve additional computational complexity, but the potential gains in model performance justify this investment.

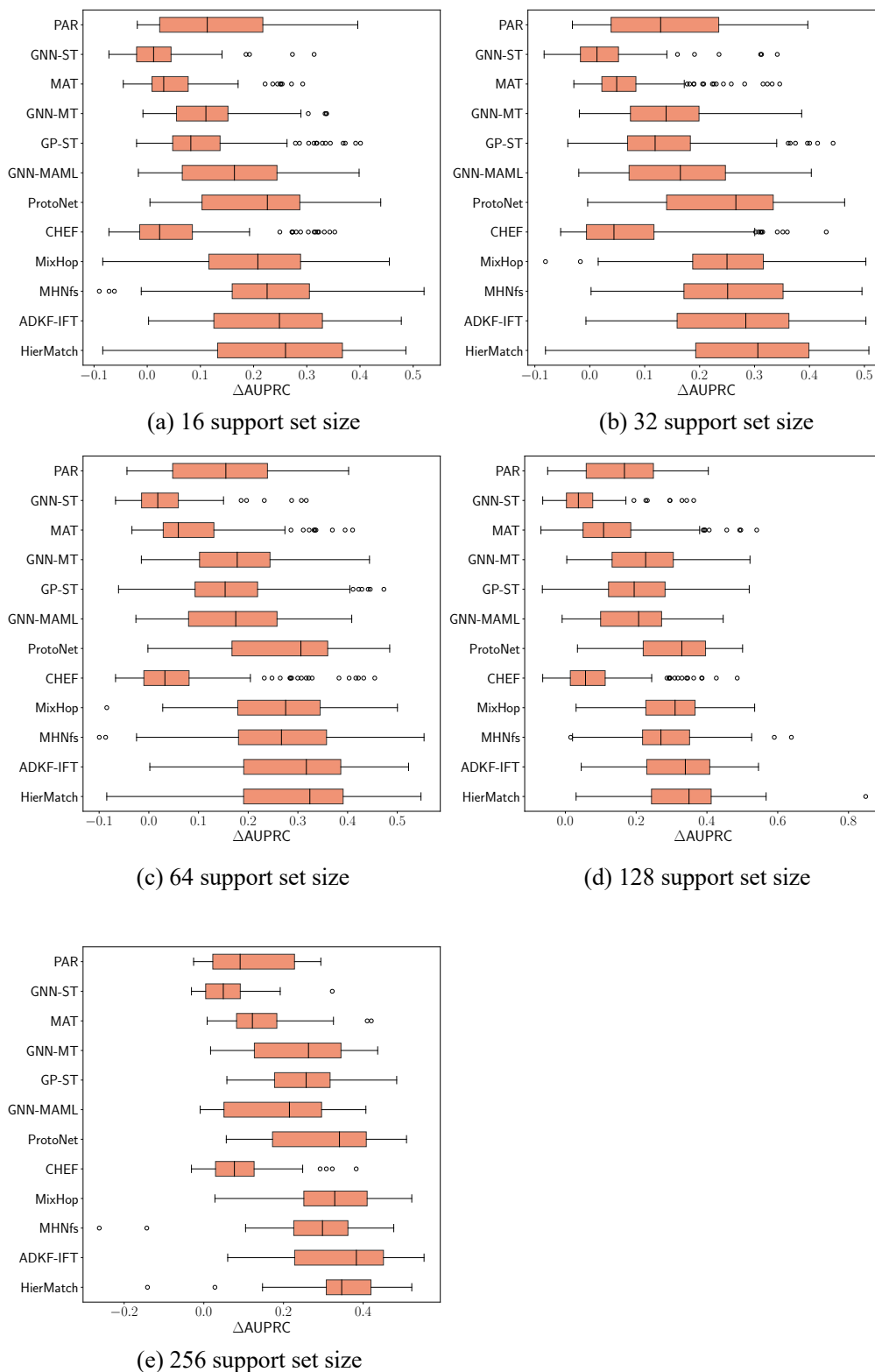


Figure 6: Box plots illustrate how different methods perform in classifying 157 FS-Mol test tasks across various support set sizes.

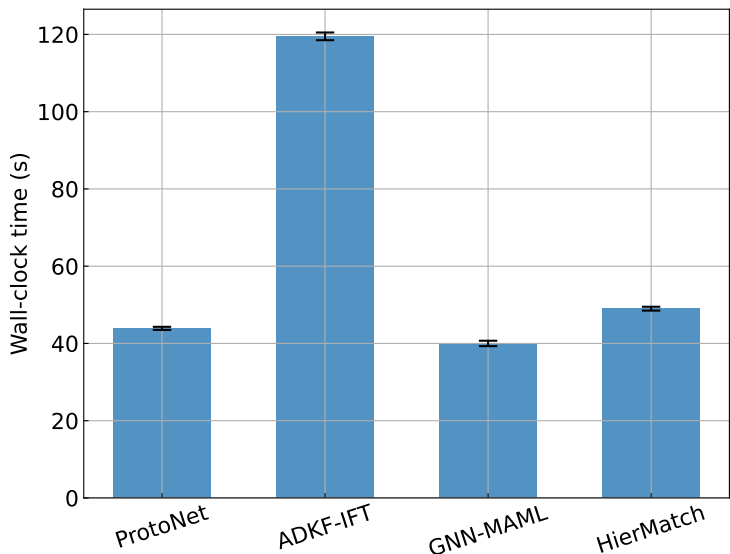


Figure 7: The wall-clock time, along with standard errors, is recorded during meta-testing on a predetermined set of FS-Mol classification tasks for comparison with the meta-learning approaches.

Limitation: Underfitting on Regression Tasks. The HierMatch model exhibits underfitting behavior on regression tasks, indicating that it may not be capturing all the necessary features and complexities required for accurate regression predictions. Experimental results suggest that this underfitting is due to the use of a linear layer in the final fusion module. When different layers’ molecular features are fused using a weighted average approach instead, the model performs significantly better and converges properly. This improvement with the weighted average fusion indicates that the linear layer may not be adequately capturing the relationships between features from different layers for regression tasks. Therefore, replacing the linear fusion with a weighted average aggregation method could resolve the underfitting issue, allowing the model to capture the necessary complexities and improve its performance on regression tasks.

Conclusion and Future Work In this paper, we propose Hierarchical Matching Networks (HierMatch) to address the limitations of existing few-shot learning approaches in drug discovery. HierMatch utilizes hierarchical pooling and matching techniques to enhance the representation and comparison of molecular features across multiple scales. By aggregating node features from local neighborhoods to global structures, HierMatch preserves essential structural details that single-scale methods often overlook. Our experimental results demonstrated that HierMatch outperforms state-of-the-art methods on the MoleculeNet and FS-Mol benchmarks, with significant improvements in AUROC and Δ AUPRC. Additionally, HierMatch showed exceptional generalization ability on the Meta-MolNet benchmark. However, our analysis revealed that the model’s performance on regression tasks could be further improved by addressing specific issues in the fusion module.

In the future, we will focus on enhancing the fusion mechanism within HierMatch to better capture the complex relationships between features from different hierarchical levels. Specifically, we will explore advanced fusion techniques such as attention-based fusion and multi-scale feature aggregation to replace the current simplistic linear approach. Additionally, we plan to conduct more extensive experiments on a wider range of datasets and tasks to ensure the robustness and generalizability of our model. Another promising direction is to integrate domain-specific knowledge and features into the model to further improve its predictive accuracy and interpretability. Finally, we will work on optimizing the computational efficiency and scalability of HierMatch to facilitate its application in large-scale drug discovery projects.