Neural Networks 163 (2023) 122–131

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Few-shot Molecular Property Prediction via Hierarchically Structured Learning on Relation Graphs

Wei Ju [a,1], Zequn Liu [a,1], Yifang Qin [b], Bin Feng [a], Chen Wang [c], Zhihui Guo [d], Xiao Luo [e,*], Ming Zhang [a,*]

[a] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, Beijing, 100871, China
[b] School of EECS, Peking University, Beijing, 100871, China
[c] College of Chemistry, Nankai University, Tianjin, 300071, China
[d] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[e] Department of Computer Science, University of California, Los Angeles, 90024, USA

ARTICLE INFO

Article history:
Received 29 September 2022
Received in revised form 25 January 2023
Accepted 22 March 2023
Available online 30 March 2023

Keywords:
Molecular property prediction
Few-shot learning
Graph neural networks
Meta learning

ABSTRACT

This paper studies few-shot molecular property prediction, which is a fundamental problem in cheminformatics and drug discovery. More recently, graph neural network based model has gradually become the theme of molecular property prediction. However, there is a natural deficiency for existing methods, that is, the scarcity of molecules with desired properties, which makes it hard to build an effective predictive model. In this paper, we propose a novel framework called Hierarchically Structured Learning on Relation Graphs (HSL-RG) for molecular property prediction, which explores the structural semantics of a molecule from both global-level and local-level granularities. Technically, we first leverage graph kernels to construct relation graphs to *globally* communicate molecular structural knowledge from neighboring molecules and then design self-supervised learning signals of structure optimization to *locally* learn transformation-invariant representations from molecules themselves. Moreover, we propose a task-adaptive meta-learning algorithm to provide meta knowledge customization for different tasks in few-shot scenarios. Experiments on multiple real-life benchmark datasets show that HSL-RG is superior to existing state-of-the-art approaches.

1. Introduction

Molecular property prediction, which aims at predicting the quantum mechanical properties of individual molecules, has been widely considered as one of the most important tasks in computational drug discovery and cheminformatics. Benefiting from the breakthrough of deep learning, this problem has raised intensive attention in recent years due to the rapid growth of available molecular structure data. It has a variety of promising applications including virtual screening and medication repurposing. Therefore, molecular property prediction plays a vital role in significantly speeding up the drug discovery process.

Actually, there are many quantitative structure property/activity relationship (QSPR/QSAR) approaches that have been proposed to achieve effective molecular property prediction. Traditionally, early works (Wang, Guo, Wang, Sun, & Huang, 2019; Zheng, Yan, Yang, & Xu, 2019) represent molecules as *SMILES* strings and leverage sequence models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to learn molecular representations. To better encode the pharmacological features of the molecules, types of fingerprint-based methods (Xu, Wang, Zhu, & Huang, 2017; Zhang et al., 2018) have been proposed for similarity comparisons for virtual screening. However, the above methods usually show the inability to model the structural properties of the molecules, since they often treat each molecule as a sequence while ignoring the intrinsic topological features.

More recently, as a molecule can be essentially represented as a graph by viewing atoms as nodes and chemical bonds as edges, graph neural networks (GNNs) (Gilmer, Schoenholz, Riley, Vinyals, & Dahl, 2017; Jiang, Chen, Wang, & Luo, 2022; Kipf & Welling, 2017; Luo, Ju, Qu, Chen, et al., 2022; Xie, Zhang, Gong, Tang, & Han, 2020; Xu, Hu, Leskovec, & Jegelka, 2019) have been widely adopted to reinforce existing molecular property prediction methods (Hao et al., 2020; Li, Zhou, Xu, Dou, & Xiong, 2022; Sun, Hoffmann, Verma, & Tang, 2020; Zhang, Liu, Wang, Lu, & Lee, 2021) via incorporating the crucial structural properties of the molecules. The basic idea of GNN-based methods is to utilize the message-passing mechanism to learn graph-level molecular representations jointly optimized with the molecular

* Corresponding authors.
E-mail addresses: xiaoluo@cs.ucla.edu (X. Luo), mzhang_cs@pku.edu.cn (M. Zhang).

[1] Equal contribution.

https://doi.org/10.1016/j.neunet.2023.03.034
0893-6080/

property prediction tasks. Specifically, ASGN (Hao et al., 2020) adopts a teacher–student framework where the teacher model learns molecular representations while the student model targets at property prediction task. MGSSL (Zhang et al., 2021) designs a general motif-based multi-level self-supervised pre-training framework in which GNNs are required to make topological and label predictions. Recently, GeomGCL (Li, Zhou, et al., 2022) leverages graph contrastive learning to capture the geometry of the molecule across 2D and 3D views.

Despite the encouraging performance achieved by GNNs, existing GNN-based models severely suffer from two key limitations: **(i) Scarcity of available molecules with desired properties**. GNN-based models inherit the characteristics of deep neural networks and are inherently data hungry, while only a small amount of labeled molecules are available to be evaluated in the lead optimization stage of drug discovery, due to a number of reasons including toxicity, low activity, and low solubility (Dahl, Jaitly, & Salakhutdinov, 2014), directly training GNNs on such limited molecules in a supervised way is prone to over-fitting and lack of generalization. **(ii) Inability to mine the molecular inherent semantic information**. The signals of supervised learning can only extract the most property-related features of the candidate molecules, while ignoring the rich structural-semantic information inherent in the molecules. The molecules themselves can serve as a regularizer, which helps a model better explore the molecular structural semantics. In view of this, the problem of label scarcity and insufficient structural semantics mining make the majority of GNN-based methods incapable of learning effective molecular representations and performing accurate property prediction. As such, we are looking for an approach that can well overcome the label scarcity, and meanwhile capture abundant semantic knowledge of the molecules.

Having realized the above challenges with existing methods, we focus on few-shot molecular property prediction to address the aforementioned limitations. Towards this end, this work proposes a principled framework called **H**ierarchically **S**tructured **L**earning on **R**elation **G**raphs (HSL-RG) for few-shot molecular property prediction. The key idea of HSL-RG is to exploit the multi-level molecular information to overcome the scarcity of laboratory molecules and insufficient structural semantics mining, and lay a solid foundation for the following wet experiment. To achieve this goal, we introduce two level objectives in the HSL-RG hierarchically, i.e., a global-level objective and a local-level objective, respectively. On the one hand, HSL-RG leverages the capability of graph kernels for capturing the structural similarity to construct relation graphs to enhance the structural knowledge communication from neighboring molecules from the global view. On the other hand, HSL-RG designs self-supervised learning signals of structure optimization to learn transformation-invariant representations from molecules themselves, endowing the molecules with desired properties from the local view. Further, the whole training process can be optimized by a novel task-adaptive meta-learning algorithm to provide meta knowledge customization for different tasks in few-shot scenarios. By incorporating this multi-level knowledge, our experiments show that it can largely improve the existing state-of-the-arts on four benchmark datasets. To summarize, the main contributions of this work are as follows:

- **General Aspects:** We propose a novel graph neural network based approach for few-shot molecular property prediction, which has been widely considered as one of the most important tasks in drug discovery.
- **Novel Methodologies:** We propose a principled framework to model the molecular structural semantics from complementary views, of which the global-level view constructs relation graphs to communicate knowledge from neighboring molecules, while the local-level view leverages self-supervised learning to achieve transformation invariance from molecules themselves. Moreover, a task-adaptive meta-learning is proposed to provide customized meta knowledge for different tasks.
- **Multifaceted Experiments:** We conduct comprehensive experiments on four benchmark datasets to demonstrate the effectiveness of the proposed approach against existing state-of-the-art models.

2. Related work

In this section, we briefly review the existing literature related to our work in four aspects, namely graph neural networks, few-shot learning, molecular property prediction, and few-shot molecular property prediction.

2.1. Graph Neural Networks (GNNs)

GNNs are originally introduced by Gori, Monfardini, and Scarselli (2005), Scarselli, Gori, Tsoi, Hagenbuchner, and Monfardini (2008), and have recently emerged as a powerful architecture to process graph-structured data, whose underlying idea is to update node representations by iteratively aggregating information from neighboring nodes via message passing (Gilmer et al., 2017; Ju, Luo, et al., 2022; Li & Cheng, 2021; Xu et al., 2019), then a readout function is applied to integrate all the node representations into a representation of the whole graph (Lee, Lee, & Kang, 2019; Rassil, Chougrad, & Zouaki, 2022; Ying et al., 2018). Molecular property prediction can be treated as a promising application of GNNs in which a molecule could be represented as a molecular graph by denoting atoms as nodes, and bonds as edges. Besides, GNNs have also shown great promise for predicting the energy (Liu, Qu, Zhang, Cai, & Tang, 2022) and other quantum mechanical properties of molecules (Luo, Ju, Qu, Gu, et al., 2022; Sun et al., 2020).

2.2. Few-shot learning

Few-shot Learning is another line of related work, which can be categorized into two main groups: (i) metric-based, and (ii) optimization-based. The former is similar to nearest neighbors and kernel density estimation, and aims to learn a metric or distance function over objects (Gao, Luo, Yang, & Zhang, 2022; Snell, Swersky, & Zemel, 2017; Sung et al., 2018; Vinyals, Blundell, Lillicrap, Wierstra, et al., 2016; Zhang, Li, & Koniusz, 2022; Zhao, Zhang, Jiang, & Tang, 2022), while the latter optimizes a meta-learner for parameter initialization which can be fast adapted to new tasks (Abbas, Xiao, Chen, Chen, & Chen, 2022; Ding et al., 2022; Finn, Abbeel, & Levine, 2017; Nichol, Achiam, & Schulman, 2018; Von Oswald et al., 2021; Ye, Wang, & Cao, 2021).

Besides, there are also some recent studies combining few-shot learning with graph neural networks (Chauhan, Nathani, & Kaul, 2020; Garcia & Bruna, 2017; Liu, Fang, Liu, & Hoi, 2021; Liu et al., 2018; Lu et al., 2022). TPN (Liu et al., 2018) learns to propagate labels between labeled instances for unlabeled test instances and is the first to model transductive inference. RALE (Liu et al., 2021) leverages the concept of hub nodes to capture the task-level and graph-level dependency based on the relative and absolute location. However, due to the complexity of domain knowledge and data structure, these methods are not tailed for molecules, and the study of few-shot learning to molecular property prediction has not been fully explored.

2.3. Molecular property prediction

Prediction of molecular properties is a central research topic in chemistry, drug discovery, and materials science. Traditional

methods therein is density functional theory (DFT) (Engel & Dreizler, 2013). However, DFT is very time-consuming and suffers from high complexity. To raise the efficiency of molecular property prediction, there has been a surge of interest in employing machine learning approaches to accelerate this process, which can be divided into two main categories based on the input molecular type: (i) simplified molecular-input line-entry (*SMILES*), and (ii) molecular graph. For the first category, *SMILES* is represented as a unique sequence that encodes the chemical species. By viewing molecules as sequences, various sequential models (e.g., RNN) in the field of natural language processing can be adopted to learn molecular representations (Wang et al., 2019; Xu et al., 2017; Zhang et al., 2018). For the second category, by viewing molecule data as graphs, GNNs have turned into the prevailing trend of learning molecular representations. A large number of approaches leverage GNNs to incorporate atom attributes and bond features (Fang et al., 2022; Hao et al., 2020; Li, Zhou, et al., 2022; Zhang et al., 2021). Nevertheless, the majority of them fail to adapt to few-shot scenarios and lack the ability to overcome the scarcity of molecules.

2.4. Few-shot molecular property prediction

There are also some recent studies for few-shot learning molecular property prediction (Guo et al., 2021; Wang, Abuduweili, Yao, & Dou, 2021). Meta-MGNN (Guo et al., 2021) designs a self-supervised module to exploit and capture unlabeled information in molecule data, and introduces a self-attentive task weight into the meta-learning framework. PAR (Wang, Abuduweili, et al., 2021) proposes a property-aware embedding function and designs an adaptive relation graph learning module to capture the relationship among molecules. However, Our proposed algorithm HSL-RG is fundamentally different from these two related works. Specifically, the proposed self-supervised module in Meta-MGNN only mines low-order structural semantic signals (bond reconstruction and atom type prediction), and fails to explore higher-order structural semantic knowledge (substructures, functional groups, and motifs), while our HSL-RG develops a novel local graph augmentation strategy called bioisosterically exchangeable replacements to achieve this goal. Additionally, different molecules in Meta-MGNN are regarded as individuals, and the relationship between them is ignored, while our approach connects these molecules via graph kernels. Compared with PAR, it captures the relationship among molecules by computing the inner product of their representations, which shows the inability to effectively mine structural knowledge. While our HSL-RG can better explore the higher-order structural similarities to capture the relationship among molecules via graph kernels. Besides, our local graph augmentation strategy can also explore the higher-order structural semantic information from the local-level view. Moreover, our developed task-adaptive meta-learning algorithm is capable of providing meta knowledge customization for different tasks in few-shot scenarios.

3. Problem definition

In this section, we give the relevant notations and formalize the problem of few-shot molecular property prediction.

Definition 1 (*Molecular Graph*)**.** A molecule can be represented as a topological graph denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_i | i = 1, \ldots, |\mathcal{G}|\}$ is the set of nodes representing chemical atoms, in which $\mathbf{x}_i$ denotes the feature vector of the node (atom) v_i indicating its type such as Carbon, Nitrogen. $\mathcal{E} = \{e_{ij} | i, j = 1, \ldots, |\mathcal{G}|\}$ is the set of edges connecting two nodes (atoms) v_i and v_j, which represent chemical bonds.

Definition 2 (*N-way K-shot*)**.** In the setting of few-shot learning, each time in the construction of the classification task, N-class data is extracted from the dataset, and each class of data is composed of K samples.

Definition 3 (*Few-shot Molecular Property Prediction*)**.** Following the setting in Meta-MGNN (Guo et al., 2021) and PAR (Wang, Abuduweili, et al., 2021), we form the few-shot problem as 2-way-K-shot molecular property prediction, in which we aim to predict whether a molecule is active or inactive on a target property, given a small number of K labeled molecules per class. Generally, we use episodic training method, which means at the training stage we sample τ-th task $\mathcal{T}_\tau \in \{\mathcal{T}_\tau\}_{\tau=1}^{N_t}$ each time, and each task contains support set $\mathcal{S}_\tau = \{(\mathcal{G}^s_{\tau,i}, y^s_{\tau,i})\}_{i=1}^{2K}$ and query set $\mathcal{Q}_\tau = \{(\mathcal{G}^q_{\tau,j}, y^q_{\tau,j})\}_{j=1}^{N^q_\tau}$, where $\mathcal{G}_{\tau,i}$ and $y_{\tau,i}$ denote molecular graph and corresponding property with index i. In essence, the objective of this problem is to learn a meta-learner from a set of tasks $\{\mathcal{T}_\tau\}_{\tau=1}^{N_t}$ and can be adapted to predict new properties with only a few labeled molecules.

4. Methodology

4.1. Overview

This paper provides a novel framework HSL-RG for few-shot molecular property prediction. At a high level, HSL-RG aims to explore the structural semantics of a molecule from both global-level and local-level. On the one hand, we leverage graph kernels to construct relation graphs to globally communicate structural knowledge from neighboring molecules. On the other hand, we design self-supervised learning of structure optimization to locally learn transformation-invariant representations from molecules themselves. To couple with two hierarchical information, a task-adaptive meta-learning is proposed to provide customized meta knowledge in few-shot scenarios. An illustration of the framework is presented in Fig. 1. Next, we first introduce the graph neural networks and the two core modules from global and local perspectives, respectively. Finally, the customized task-adaptive meta-learning algorithm is explained.

4.2. Graph Neural Networks (GNNs)

GNNs (Gilmer et al., 2017) treat the molecule as a graph, in which the node is a chemical atom and the edge is a chemical bond between two atoms. Specifically, GNNs consist of L message passing layers. At lth layer, the representation of node v is updated by iteratively aggregating node representations of its neighbors $\mathcal{N}(v)$ by passing messages along the edges and the representation of node v itself. Formally, the updating process can be defined as follows:

$$\mathbf{h}_v^{(l)} = \mathcal{C}^{(l)}\left(\mathbf{h}_v^{(l-1)}, \mathcal{A}^{(l)}\left(\left\{\mathbf{h}_v^{(l-1)}, \mathbf{h}_u^{(l-1)}, e_{vu}\right\}_{u\in\mathcal{N}(v)}\right)\right), \tag{1}$$

where $\mathbf{h}_v^{(l)}$ denotes the representation of node v at layer l, e_{vu} denotes edge feature between nodes v, u. Here $\mathcal{A}^{(l)}$ and $\mathcal{C}^{(l)}$ denote the aggregation and combination functions at layer l. After L iterations, we can take the average or sum operations to integrate all node representations to obtain the whole molecular representation $\mathbf{h}_\mathcal{G}$ via READOUT function:

$$\mathbf{h}_\mathcal{G} = \text{READOUT}(\{\mathbf{h}_v^L : v \in \mathcal{V}\}), \tag{2}$$

Pre-trained Graph Neural Network. Different from the random initialization of atoms and bonds in the molecular graph, motivated by the prominent success of pre-training which has shown to be effective in many language (Devlin, Chang, Lee,

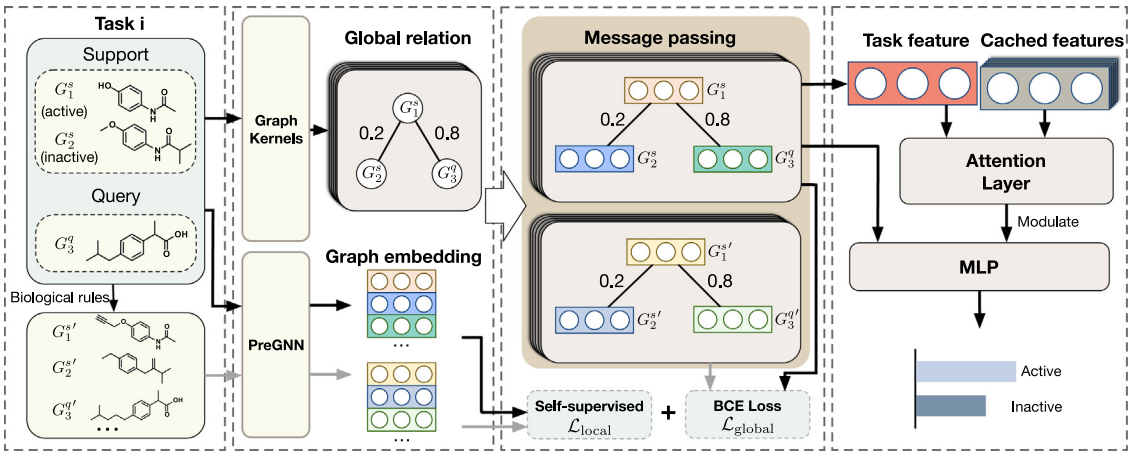


Fig. 1. Illustration of the proposed framework HSL-RG. The black arrows represent the propagation of the original graphs and the gray arrows represent the propagation of the augmented graphs. First, HSL-RG leverages graph kernels to construct relation graphs from the global view. Then, HSL-RG designs self-supervised learning of structure optimization from the local view. Finally, a task-adaptive meta-learning is proposed to provide meta knowledge customization for different tasks in few-shot scenarios.

& Toutanova, 2018), vision (He, Zhang, Ren, & Sun, 2016) and graph domains (Hu, Dong, Wang, Chang, & Sun, 2020; Qiu et al., 2020), we adopt the recent pre-trained graph neural network (PreGNN) (Hu et al., 2019) to provide a better parameter initialization. In this way, the initial features of atoms and bonds can well capture the universal information shared by molecules. Subsequent experiments have provided further evidence supporting the necessity of this approach.

4.3. Global relation graph construction

There is a fundamental assertion that similar molecules will tend to exhibit similar properties, which can either be physical or chemical. For example, hexane and heptane should have similar boiling points and water solubility with the same chain-like structure, cocaine and procaine are both local anesthetics sharing the same functional groups.

Nevertheless, existing methods solely treat molecules as individual instances in the training stage, which are independent of each other, and they hence fail to derive extra supervision signals from other molecules. To this end, we draw inspiration from that molecular structure often determines its properties, we propose to leverage graph kernels to construct a global relation graph connecting similar graphs to incorporate an additional data source, thus transferring prior knowledge and guiding the optimization of the molecules.

Technically, graph kernels (GKs) (Gärtner, Flach, & Wrobel, 2003; Kashima, Tsuda, & Inokuchi, 2003) have shown great superiority in capturing high-order substructures (e.g., random walk (Gärtner et al., 2003), path (Kashima et al., 2003), motif (Shervashidze, Vishwanathan, Petri, Mehlhorn, & Borgwardt, 2009), subtree (Shervashidze, Schweitzer, Van Leeuwen, Mehlhorn, & Borgwardt, 2011)). In view of this, we propose to leverage random walk graph kernels (Gärtner et al., 2003; Ju, Yang, et al., 2022) to capture the functional groups in molecules, due to its capability of exploring the motif patterns which is critical in biochemistry (Wang, Guo, Ju, Luo, & Deng, 2021). Here we give the clear definition of graph kernels.

Definition 3: (Graph Kernels) Given two graphs $G = (\mathcal{V}, \mathcal{E})$ and $G' = (\mathcal{V}', \mathcal{E}')$, the graph kernel $Ker(G, G')$ measures the similarity between them and is defined as:

$$Ker\left(G, G'\right) = \sum_{v \in \mathcal{V}} \sum_{v' \in \mathcal{V}'} k_{base}\left(f_G\left(v\right), f_{G'}\left(v'\right)\right), \tag{3}$$

where base kernel k_{base}, i.e., inner product on Hilbert space, compares substructures $f_G(\cdot)$ centered at nodes v and v', in which $f_G(\cdot)$ denote the feature vector counting the number of frequencies of each substructure (e.g., graphlets, random walks, paths, subtrees) in the graph G.

Since GKs inherently involve similarity comparisons between substructure patterns, we hence utilize this characteristic to construct a KNN graph as the global relation graph, which connects these individual molecules in each task. Specifically, the KNN graph is calculated by the similarity matrix $\mathbf{S} \in \mathbb{R}^{(2K+1)\times(2K+1)}$ of the $2K$ support molecules and 1 query molecule. Each entry $\mathbf{S}_{ij}$ measures the structural similarity between two molecules G_i and G_j computed by $\mathbf{S}_{i,j} = Ker\left(G_i, G_j\right)$ if G_j is in the most K similar neighbors of G_i, otherwise $\mathbf{S}_{ij}$ is set to 0.

In this way, the global relation graph provides prior knowledge on the structural similarity between different molecules, which allows us to transfer supervision between these molecules and even generalize to unseen molecules.

Message Passing on Global Relation Graph. After the construction of the global relation graph, we can leverage message passing neural networks to propagate molecular messages. Here we use GIN (Xu et al., 2019) to learn node representations (each molecule) on global relation graph through Eq. (1) denoted as $\mathbf{z}^l = \text{GIN}(\mathbf{S}, \mathbf{z}^{(l-1)})$, in which $\mathbf{z}^{(0)}$ represents the initial feature of each individual molecule $\mathbf{h}_{\mathcal{G}}$ derived from previously PreGNN (Hu et al., 2019). After several iterations of message passing, the molecules on the global relation graph can communicate with each other and propagate molecular structural knowledge to neighboring molecules. The learned molecular representations $\mathbf{z}_{\tau,i}$ can absorb property-related features from neighbors, enriching themselves for better molecular property prediction.

Formally, a downstream classifier (i.e., multi-layer perceptron, MLP) is applied to predict the probability of target property of each molecule $\mathbf{z}_{\tau,i}$: $\hat{y}_{\tau,i} = \text{MLP}(\mathbf{z}_{\tau,i})$. Hence the global-level objective is evaluated via binary cross-entropy loss defined as:

$$\mathcal{L}_{\text{global}} = -\sum_{(\mathcal{G}^S_{\tau,i}, y^S_{\tau,i}) \in \mathcal{S}_\tau} [y_{\tau,i} \log \hat{y}_{\tau,i} + (1 - y_{\tau,i}) \log(1 - \hat{y}_{\tau,i})], \tag{4}$$

where $y_{\tau,i}$ is the ground-truth target property.

4.4. Local graph augmentation optimization

Though incorporating molecular prior knowledge via the global relation graph could derive extra supervision signals of properties, the performance is still far from satisfactory due to the

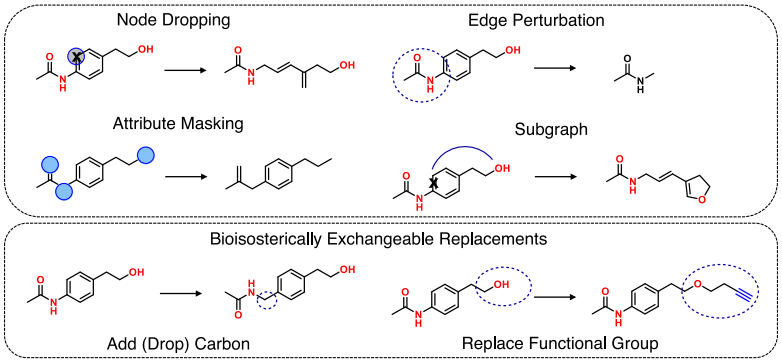


Fig. 2. Illustration of the augmentation comparison. Upper: traditional augmentations may alter the molecular semantics. Lower: the augmentation via bioisosterically exchangeable replacements could peserve inherent properties.

limited molecules of each task in few-shot scenarios. Inspired by the success of self-supervised learning which possesses its powerful capability in learning effective representations from molecules themselves, we aim to leverage this technique to further overcome the scarcity of label properties and fully capture structural semantics.

The basic idea of self-supervised learning is to transform the data to generate augmented views and thus can learn representations that are invariant to transformations by achieving consistency between augmented views. Specifically, GraphCL (You et al., 2020) introduces four types of augmentations for general graphs shown in the upper of Fig. 2. However, these heuristic augmentations are prone to destroy the semantics of molecules. For example, the original molecule has certain hydrophilicity, while edge perturbation might introduce an epoxide, which makes the molecule more hydrophobic, drastically changing the properties.

To this end, we introduce a new augmentation strategy called bioisosterically exchangeable replacements, in which a valid substructure in a molecule is replaced by a bioisostere (Meanwell, 2011). By doing so, this strategy would produce a new molecule with similar physical or chemical properties as the original one. Technically, we use BoBER dataset[2] (Lešnik et al., 2017) to build our rule set. BoBER consists of molecular fragment pairs that have similar binding properties, which means that they are found to bind to similar binding sites. BoBER is constructed by mining the Protein Data Bank (PDB) using the ProBiS algorithm (Konc & Janežič, 2010), and contains 14 407 similar fragment pairs. However, it is nontrivial to directly utilize them for augmentation, and most fragment pairs have low evidence (the number of similar binding sites in the PDB that bind to both fragments), replacements built on them might lead to unfaithful bioisostere results. Therefore, we first rank all fragment pairs in BoBER by their evidence, and finally select 224 pairs to build our rule set.

Our rules are formulated by SMARTS strings, which are commonly used to describe molecular patterns and are broadly used in substructure searching and chemical reaction representation. As bioisostere replacement can also be seen as a chemical reaction, we hence leverage SMARTS to represent our replacement rules. To build our rule set, we first use ChemDraw[3] to draw our selected molecular fragment pairs, and then we align atom pairs manually, and finally export rules represented by SMARTS strings. A sample rule is as follows:

$$[*:1][C:2](=O)O[c,C:3] \gg [*:1][C:2](=O)N(C)[*:3]$$

[2] http://insilab.org/datasets/

[3] https://chemdrawdirect.perkinelmer.cloud/js/sample/index.html

Algorithm 1 Training algorithm of HSL-RG

Input: Support set $\mathcal{S}_\tau$, query set $\mathcal{Q}_\tau$
Output: Initialized parameters θ, task-adaptive gate parameters ϕ and task features $\mathbf{S}_{train}$

$\mathbf{S}_{train} \leftarrow \mathbf{0}$.
while not done **do**
  sample a batch of tasks $\mathcal{T}_\tau$.
  For all $\mathcal{T}_\tau$ **do**
  sample support set $\mathcal{S}_\tau$ and query set $\mathcal{Q}_\tau$ from $\mathcal{T}_\tau$.
  calculate the task feature $\mathbf{s}_\tau$ and modulate θ by Eqs. (6)–(8).
  update task features $\mathbf{S}_{train}[\tau] \leftarrow \mathbf{S}_{train}[\tau] + \mathbf{s}_\tau$
  update θ' by Eq. (9).
  end for
  update θ and ϕ by Eq. (11).
end while

In this way, we generate T augmented molecules for each molecule G_i with our augmentation strategy. Then we define the local-level objective by achieving the consistency between each augmented molecule and their average representation for self-supervised learning:

$$\mathcal{L}_i^{\text{local}} = \frac{1}{T}\sum_{t=1}^{T}\left(\|\widetilde{\mathbf{h}}_i^t - \frac{1}{T}\sum_{k=1}^{T}\widetilde{\mathbf{h}}_i^k\|_2 + \|\widetilde{\mathbf{z}}_i^t - \frac{1}{T}\sum_{k=1}^{T}\widetilde{\mathbf{z}}_i^k\|_2\right)$$
$$\mathcal{L}_{\text{local}} = \sum_{(\mathcal{G}_{\tau,i}^S, y_{\tau,i}^S)\in\mathcal{S}_\tau} \mathcal{L}_i^{\text{local}}, \tag{5}$$

where $\widetilde{\mathbf{h}}_i^t$ and $\widetilde{\mathbf{z}}_i^t$ are the representations of the tth augmentation of G_i before and after the message passing on global relation graph, respectively.

4.5. Task-adaptive meta-learning

To well adapt our approach to the few-shot scenarios, we resort to the meta-learning framework based on MAML (Finn et al., 2017). In meta-training stage, it require a meta-learner to learn a good initialized parameter $\theta = (\theta_e, \theta_g, \theta_c)$, where θ_e, θ_g and θ_c are the parameters for the PreGNN, the global graph neural network, and classifier respectively. Considering that some tasks are similar and should share similar classifiers, we introduce a task-adaptive gate parameterized by ϕ to explicitly modulate θ_c with the task feature and task relationship. Specifically, we represent a task τ with the prototype of its samples that are active and those that are inactive:

$$\mathbf{s}_\tau = \text{Mean}(\{\mathbf{z}_i^+\}_{i=1}^K) \parallel \text{Mean}(\{\mathbf{z}_i^-\}_{i=1}^K), \tag{6}$$

where $\{\mathbf{z}_i^+\}_{i=1}^K$ are the active samples and $\{\mathbf{z}_i^-\}_{i=1}^K$ are the inactive samples in task τ. Mean operation denotes averaging all embeddings. $\|$ is the concatenation operator.

To encode the relationship between tasks, we combine the task feature $\mathbf{s}_\tau \in R^d$ with cached task features of all the N_t meta-training tasks by attention mechanism:

$$\mathbf{s}'_\tau = \text{Attention}(\mathbf{S}_\tau, \mathbf{S}_{train}, \mathbf{S}_{train}), \tag{7}$$

where $\mathbf{S}_\tau \in R^{1\times d}$ is the matrix form of $\mathbf{s}_\tau$, representing a task feature for a meta-training task, that is, a row in $\mathbf{S}_{train} \in R^{N_t\times d}$. Attention used here is the multi-head attention mechanism in Vaswani et al. (2017).

Then we learn a task-specific gate β parameterized by the task feature s'_τ and further use it to modulate θ_c:

$$\boldsymbol{\beta} = f(\mathbf{s}'_\tau), \boldsymbol{\theta}_c = \boldsymbol{\theta}_c \circ \boldsymbol{\beta}, \tag{8}$$

where f is a MLP and $\circ$ is the element-wise multiplication.

Following Lu et al. (2019), we fix θ_e, θ_g, ϕ, and update θ_c in each τ-th task $\mathcal{T}_\tau$ through a small number of gradient descent with support set:

$$\theta'_{c,\tau} = \theta_c - \alpha\nabla_\theta \mathcal{L}_{\mathcal{T}_\tau}(\theta, \phi). \tag{9}$$

where α is the learning rate of the adaptation. $\mathcal{L}_{\mathcal{T}_\tau}(\cdot)$ consists of two component: global-level and local-level objectives, which proceed as described in Eqs. (4) and (5) respectively. Therefore, the overall training objective for each task can be written as (λ is tuning parameter):

$$\mathcal{L}_{\mathcal{T}_\tau}(\cdot) = \mathcal{L}_{\text{global}} + \lambda\mathcal{L}_{\text{local}}, \tag{10}$$

Subsequently, the meta-objective can be optimized via integrating the training objectives of all sampled tasks:

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \sum_{\tau=1}^{N_t} \mathcal{L}'_{\mathcal{T}_\tau}(\theta'_\tau, \phi), \tag{11}$$

where $\mathcal{L}'_{\mathcal{T}_\tau}$ is the joint loss over query set $\mathcal{Q}_\tau$.

During meta-testing, the meta-learner has collected transferable knowledge and would be further adapted to support set $\mathcal{S}_{\text{new}}$ of each new task $\mathcal{T}_{\text{new}}$ and evaluated on the query set $\mathcal{Q}_{\text{new}}$. The overall framework is shown in Algorithm 1.

4.6. Computational complexity analysis

Suppose that N_q is the average size of query set in each task, $|\mathcal{E}|$ is the average size of edge set of each molecule, the complexity of GIN (Xu et al., 2019) for each molecular is $O(|\mathcal{E}|)$, and the total complexity in this stage is $O(N_t(2K + N_q)|\mathcal{E}|) \approx O(N_t N_q|\mathcal{E}|)$. Similarly, the global message passing can be done within $O(N_t N_q)$. Collectively, the total computational complexity is $O(N_t N_q|\mathcal{E}|)$.

5. Experiment

In this section, we conduct extensive experiments on four molecular property prediction datasets. We attempt to answer the following research questions:

- **RQ1:** How well does our HSL-RG perform against the baseline models on molecular property prediction?
- **RQ2:** How does each part of the model affect the molecular property prediction? How do hyper-parameters influence the model performance?
- **RQ3:** How can we intuitively show the effectiveness of the global relation graph and the attention mechanism in task-adaptive meta-learning?

Table 1
Data statistics.

Dataset	Tox21	SIDER	MUV	ToxCast
# Tasks	12	27	17	617
# Training Tasks	9	21	12	450
# Testing Tasks	3	6	5	167
# Compounds per Task	667.83	52.85	5478.06	13.96

5.1. Experimental setup

Datasets. We evaluate our method on four molecular property prediction benchmark datasets from MoleculeNet (Wu et al., 2018), a wide standard benchmark which has 17 types of molecular property in total.

- **Tox21** (National Center for Advancing Translational Sciences, 2017) aims to predict the human toxicity of 8014 compounds on 12 different targets.
- **SIDER** (Kuhn, Letunic, Jensen, & Bork, 2016) contains 1427 compounds used in marketed medicines with 27 categories of side effects.
- **MUV** (Rohrer & Baumann, 2009) contains 93 127 compounds for validating virtual screening.
- **ToxCast** (Richard et al., 2016) comprises 8615 chemical compounds with toxicity labels that are obtained through high-throughput screening.

We follow the data splits in Wang, Abuduweili, et al. (2021). Table 1 shows the detailed statistics of benchmark datasets. Given the small number of compounds in each task, training models for each task from scratch is hard, necessitating few-shot molecular property prediction.

Baselines. We compare our method with types of methods for few-shot molecular property prediction. We consider the Few-shot learning (FSL) methods learned from scratch:

- **Siamese** (Koch, Zemel, Salakhutdinov, et al., 2015) leverages dual convolutional neural networks and matches the query sample to the support samples.
- **ProtoNet** (Snell et al., 2017) learns class prototypes with the support set and calculates the distances between the query sample and the prototypes.
- **IterRefLSTM** (Altae-Tran, Ramsundar, Pappu, & Pande, 2017) utilizes the idea of matching network for few-shot molecular property prediction.
- **MAML** (Finn et al., 2017) trains a meta-learner to learn parameter initialization.
- **TPN** (Liu et al., 2018) constructs a relation graph and propagate the node labels.
- **EGNN** (Kim, Kim, Kim, & Yoo, 2019) predicts edge labels of the relation graph.
- **Sharp-MAML** (Abbas et al., 2022) is a sharpness-aware MAML method which avoids the sharp local minima of MAML loss functions.

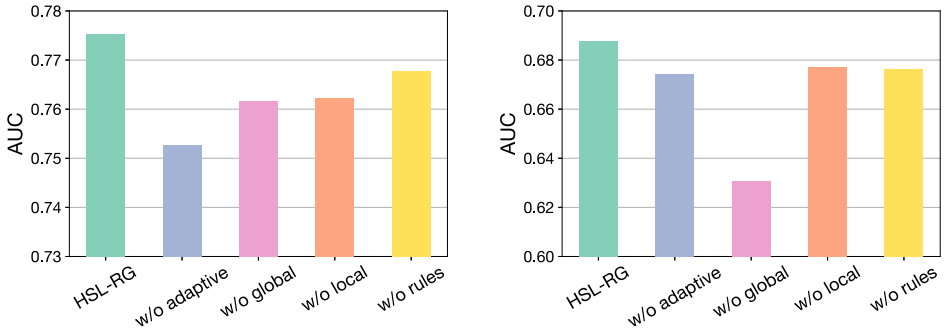
We also compared the variant of our model HSL-RG$^-$ which is trained from scratch without the initialization of Pre-GNN.

We also adopt the pretraining-based methods:

- **Pre-GNN** (Hu et al., 2019) pre-trains a GIN (Xu et al., 2019) with self-supervised tasks.
- **Meta-MGNN** (Guo et al., 2021) is a MAML-based method using the pretrained parameters of Pre-GNN as the initialization of the molecular encoder.
- **Pre-PAR** (Wang, Abuduweili, et al., 2021) learns an adaptive relation graph among molecules for each task and also uses Pre-GNN to initialize the molecular encoder.

Table 2
AUC on benchmark molecular property prediction datasets.

Method	Tox21		SIDER		MUV		ToxCast	
	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot
Siamese	80.40 ± 0.35	65.00 ± 1.58	71.10 ± 4.32	51.43 ± 3.31	59.96 ± 5.13	50.00 ± 0.17	–	–
ProtoNet	74.98 ± 0.32	65.58 ± 1.72	64.54 ± 0.89	57.50 ± 2.34	65.88 ± 4.11	58.31 ± 3.18	63.70 ± 1.26	56.36 ± 1.54
MAML	80.21 ± 0.24	75.74 ± 0.48	70.43 ± 0.76	67.81 ± 1.12	63.90 ± 2.28	60.51 ± 3.12	66.79 ± 0.85	65.97 ± 5.04
TPN	76.05 ± 0.24	60.16 ± 1.18	67.84 ± 0.95	62.90 ± 1.38	65.22 ± 5.82	50.00 ± 0.51	62.74 ± 1.45	50.01 ± 0.05
EGNN	81.21 ± 0.16	79.44 ± 0.22	72.87 ± 0.73	70.79 ± 0.95	65.20 ± 2.08	62.18 ± 1.76	63.65 ± 1.57	61.02 ± 1.94
IterRefLSTM	81.10 ± 0.17	80.97 ± 0.10	69.63 ± 0.31	71.73 ± 0.14	49.56 ± 5.12	48.54 ± 3.12	–	–
Sharp-MAML	75.37 ± 0.23	74.59 ± 0.56	71.02 ± 0.81	68.43 ± 0.96	65.52 ± 2.01	65.12 ± 2.98	67.56 ± 1.01	66.49 ± 1.98
HSL-RG$^-$	80.95 ± 0.26	79.65 ± 0.22	74.66 ± 0.52	71.77 ± 0.79	70.38 ± 1.35	67.22 ± 1.56	70.70 ± 1.02	70.06 ± 1.05
Pre-GNN	82.14 ± 0.08	81.68 ± 0.09	73.96 ± 0.08	73.24 ± 0.12	67.14 ± 1.58	64.51 ± 1.45	73.68 ± 0.74	72.90 ± 0.84
Meta-MGNN	82.97 ± 0.10	82.13 ± 0.13	75.43 ± 0.21	73.36 ± 0.32	68.99 ± 1.84	65.54 ± 2.13	–	–
Pre-PAR	84.93 ± 0.11	83.01 ± 0.09	78.08 ± 0.16	74.46 ± 0.29	69.96 ± 1.37	66.94 ± 1.12	75.12 ± 0.84	73.63 ± 1.00
HSL-RG	**85.56 ± 0.28**	**84.09 ± 0.20**	**78.99 ± 0.33**	**77.53 ± 0.41**	**71.26 ± 1.08**	**68.76 ± 1.05**	**76.00 ± 0.81**	**74.40 ± 0.82**



(a) SIDER (b) MUV

Fig. 3. Ablation study. Results of different variants of HSL-RG on 1-shot setting of **(a)** SIDER and **(b)** MUV.

Evaluation Metrics. We adopt the average **AUC** on testing tasks to evaluate the molecular property prediction performance. We run experiments on all the datasets for five times and report the mean and standard deviations. AUC is the area under ROC curves (Receiver Operating Characteristic curves), which is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive Rate.

Implementation Details. We use a 2-layer GIN to model the global relational graph, the hidden size is set to 128. We use a 2-layer MLP on top of the global graph embedding for classification. N is set to 2 and λ is set to 0.1. We follow the other hyper-parameter settings in Wang, Abuduweili, et al. (2021). When meta-testing, each time we select one sample from the query set and adapt the model on the support set. We repeat the process for N_q times in total. N_q is the size of the query set for each task. All experiments are carried out on NVIDIA GeForce RTX 3090. We use one GPU and the whole training and evaluation process can be finished within four hours.

5.2. Overall performance comparison

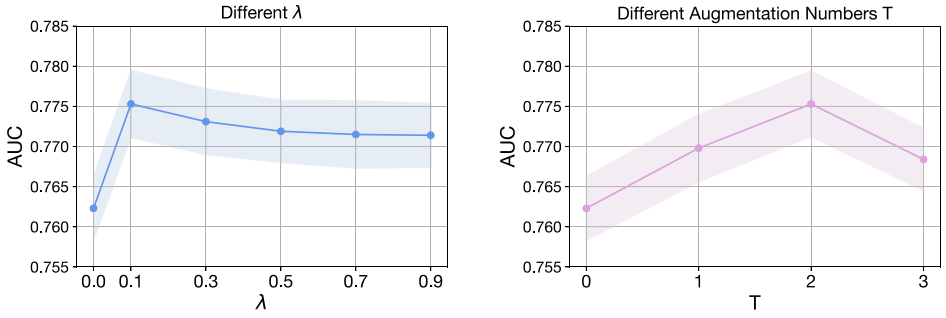
Here, we evaluate the performance of all the algorithms and the results are summarized in Table 2. From the comprehensive views, we have several observations:

- Our HSL-RG shows superior performance and achieves the best results across all four datasets on 1-shot and 10-shot settings. For example, HSL-RG obtains more than 4.1% AUC improvement against the best baseline Pre-PAR on the 1-shot setting of MUV.
- Approaches enhanced by the pre-trained model perform substantially better than methods learned from scratch, indicating the effectiveness of the pre-trained model in few-shot molecular property prediction. For instance, Pre-GNN achieves better performance than the few-shot learning methods trained from scratch despite their sophisticated design for few-shot settings.
- The relation between molecules benefits the property predictions. Our proposed method and the baselines EGNN and Pre-PAR considering the relation graph obtain relatively better performance than other baselines. For example, the AUCs of Pre-PAR on all the datasets are higher than Meta-MGNN although they have the same PreGNN encoder and MAML framework.

5.3. Ablation and hyper-parameter study

To further study the contribution of each component of the proposed HSL-RG, we conduct the ablation study and hyperparameter study, which reveals the actual mechanism behind the whole process of predicting molecular properties. **Ablation study.** We implement four variants of our model: (1) **w/o local** is the variant of our model without local graph augmentation. (2) **w/o global** is the variant of our model without message passing on the global relation graph. (3) **w/o rules** is the variant of our model using random augmentations instead of rule-based augmentation. (4) **w/o adaptive** is the variant of our model without the task adaptive gate. We evaluate these variants on the 1-shot setting of SIDER and MUV as shown in Fig. 3. We find that:

- The performance drops on w/o local and w/o rules, verifying the effectiveness of our proposed local graph augmentation. w/o rules performs better than w/o local on both datasets, showing that traditional augmentation methods slightly improve the graph embedding.



(a) Weight λ (b) Number T

Fig. 4. Hyperparameter study on the 1-shot setting of SIDER. **(a)** Effects of the weight of the local-level objective loss λ and **(b)** the augmentation number T.

- The global relation graph is essential to our proposed model, since the w/o global results decrease significantly compared to HSL-RG. This provides further evidence that the global relation graph enriches the molecular embeddings, leading to improved predictions.
- w/o adaptive has less superior performance than HSL-RG, demonstrating the effectiveness of the task-adaptive gate. However, it still outperforms baselines, reassuring the advantage of the hierarchical graph.

Hyper-parameter Study. On the 1-shot setting of SIDER, we study the sensitivity of hyper-parameters in the local graph augmentation module, that is, the weight of the local-level objective loss λ, and the augmentation number T. Fig. 4 shows the results. It can be observed that:

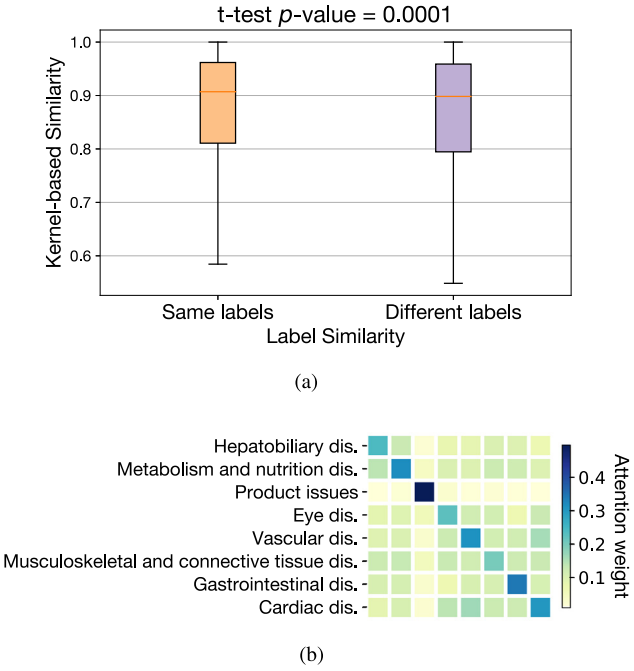
- The local graph augmentation module plays an essential role in HSL-RG. When T or λ is set to 0, the model would degenerate into the w/o local variant and the performance has a substantial degradation.
- A proper weight for the local-level objective loss and a proper augmentation number could assist the global and local GNNs in better learning the representations of molecules and predicting molecular properties. The model performs the best when λ is set to 0.1 and T is set to 0, demonstrating the correctness of our hyper-parameter selection based on the evaluation of the validation set.
- The stability of the model is satisfactory. When T or λ increases, the AUC of our model increases first and then drops slightly but is still not less than 0.76, which is significantly higher than the best baseline.

5.4. Visualization

We intuitively validate the superiority of the global relation graph using graph kernel and the attention-based task-adaptive mechanism by visualization.

Correlation between substructure similarity and property similarity. The hypothesis of our framework is based on that molecules with similar substructures tend to have similar properties. We validate this hypothesis on ToxCast in the experiment. Specifically, We randomly sample 10 000 pairs of molecules. For each pair, we calculate their kernel-based similarity scores, and examine whether they have the same labels on "ACEA_T47D_80h_Negative". We observe a substantial agreement with 0.0001 p-value (Fig. 5a) between the kernel-based similarity and label similarity, indicating the correctness and superiority to construct a global relation graph via our graph kernels.

Visualization of the attention weights in the task adaptive gate. The attention weights in the task adaptive module can



(a)

(b)

Fig. 5. Visualization experiments. **(a)** Correlation between kernel-based similarity and label similarity. **(b)** Visualization of the attention weights among eight tasks on SIDER. (dis. is short for the disorder).

reflect the task relationship. Specifically, higher attention weights show stronger relevance. We select eight tasks in SIDER and calculate their attention weights (Fig. 5b). Each task represents a category of side effects. We can see that the results are in line with common sense in general. For example, the "product issues" are not relevant to other tasks representing disorders and the attention weights are low, the "cardiac disorder" has a stronger correlation with "vascular disorder" than other tasks and the attention weights between them are relatively higher.

5.5. Case study

We study a case on the 1-shot setting of SIDER to explain the prediction process of HSL-RG in depth. Fig. 6 shows three molecules in SIDER together with their SMILES and KNN relation graph calculated by graph kernels. The task is to predict whether the query molecule is active on the property "Reproductive system and breast disorders". The query molecule is connected to the active support molecule with a high edge weight of 0.69. After the message passing on the global relation graph, the model correctly

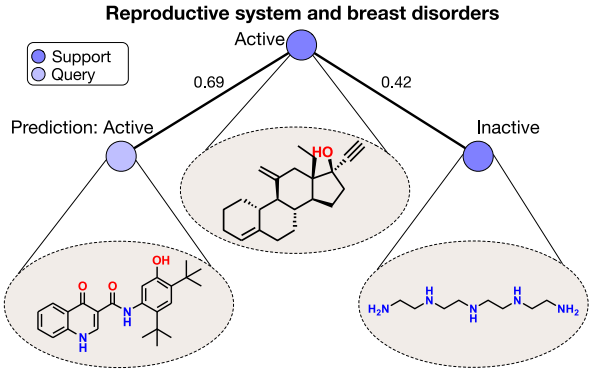


Fig. 6. Case study of the prediction of "reproductive system and breast disorders" on the 1-shot setting of SIDER.

predicts its label "active" with the help of this edge. These results can indicate the effectiveness of the global relation graph.

6. Broader impact

Our proposed algorithm HSL-RG is built on the setting of few-shot learning, however, it can be well generalized to the case of zero-shot learning, which aims at classifying samples from unseen classes that have never appeared in the training data. Zero-shot learning (Xie, Zhang, Xiong, Shao, & Li, 2022) has attracted great attention in a range of applications, such as image classification (Li, Yang, Wei, Deng, & Yang, 2022), object recognition (Zablocki, Bordes, Soulier, Piwowarski, & Gallinari, 2019) and knowledge graph completion (Geng et al., 2022). By incorporating class semantic knowledge and capturing the relations between all classes, our approach can well extend to zero-shot learning, and transfer knowledge from seen classes to unseen classes with the guidance of some auxiliary semantic information.

7. Conclusion

In this paper, we introduce a novel framework called Hierarchically Structured Learning on Relation Graphs (HSL-RG) for few-shot molecular property prediction, which explores the structural semantics of a molecule from both global-level and local-level granularities. We first leverage graph kernels to construct global relation graphs from neighboring molecules while then designing self-supervised learning of local structure optimization from molecules themselves. Moreover, a task-adaptive meta-learning algorithm is proposed to provide customized meta knowledge. Experiments well demonstrate the superiority of our method on a variety of benchmarks. Our future works will further extend our framework to other domains such as drug–drug interaction prediction, drug binding structure prediction and molecular conformation generation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors are grateful to the anonymous reviewers for critically reading the manuscript and for giving important suggestions to improve their paper.

This paper is partially supported by grants from the National Key Research and Development Program of China with Grant No. 2018AAA0101902 as well as the National Natural Science Foundation of China (NSFC Grant No. 62276002).

References

Abbas, M., Xiao, Q., Chen, L., Chen, P.-Y., & Chen, T. (2022). Sharp-MAML: Sharpness-aware model-agnostic meta learning. In *International conference on machine learning*.

Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Central Science*, *3*(4), 283–293.

Chauhan, J., Nathani, D., & Kaul, M. (2020). Few-shot learning on graphs via super-classes based on graph spectral measures. arXiv preprint arXiv:2002.12815.

Dahl, G. E., Jaitly, N., & Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. arXiv preprint arXiv:1406.1231.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, Y., Wu, Y., Huang, C., Tang, S., Yang, Y., Wei, L., et al. (2022). Learning to learn by jointly optimizing neural architecture and weights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 129–138).

Engel, E., & Dreizler, R. M. (2013). *Density functional theory*. Springer.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., et al. (2022). Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, *4*(2), 127–134.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126–1135). PMLR.

Gao, F., Luo, X., Yang, Z., & Zhang, Q. (2022). Label smoothing and task-adaptive loss function based on prototype network for few-shot learning. *Neural Networks*.

Garcia, V., & Bruna, J. (2017). Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Learning theory and kernel machines* (pp. 129–143). Springer.

Geng, Y., Chen, J., Zhang, W., Xu, Y., Chen, Z., Pan, J. Z., et al. (2022). Disentangled ontology embedding for zero-shot learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 443–453).

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

Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks (vol. 2), no. 2005* (pp. 729–734).

Guo, Z., Zhang, C., Yu, W., Herr, J., Wiest, O., Jiang, M., et al. (2021). Few-shot graph learning for molecular property prediction. In *Proceedings of the web conference 2021* (pp. 2559–2567).

Hao, Z., Lu, C., Huang, Z., Wang, H., Hu, Z., Liu, Q., et al. (2020). ASGN: An active semi-supervised graph neural network for molecular property prediction. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 731–752).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, Z., Dong, Y., Wang, K., Chang, K.-W., & Sun, Y. (2020). Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1857–1867).

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., et al. (2019). Strategies for pre-training graph neural networks. arXiv preprint arXiv:1905.12265.

Jiang, B., Chen, S., Wang, B., & Luo, B. (2022). MGLNN: Semi-supervised learning via multiple graph cooperative learning neural networks. *Neural Networks*.

Ju, W., Luo, X., Ma, Z., Yang, J., Deng, M., & Zhang, M. (2022). GHNN: Graph Harmonic Neural Networks for semi-supervised graph-level classification. *Neural Networks*, *151*, 70–79.

Ju, W., Yang, J., Qu, M., Song, W., Shen, J., & Zhang, M. (2022). KGNN: Harnessing kernel-based networks for semi-supervised graph classification. In *Proceedings of the fifteenth ACM international conference on web search and data mining* (pp. 421–429).

Kashima, H., Tsuda, K., & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. In *Proceedings of the 20th international conference on machine learning* (pp. 321–328).
Kim, J., Kim, T., Kim, S., & Yoo, C. D. (2019). 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).
Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.
Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop (vol. 2)*. Lille.
Konc, J., & Janežič, D. (2010). ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, *26*(9), 1160–1168.
Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research*, *44*(D1), D1075–D1079.
Lee, J., Lee, I., & Kang, J. (2019). Self-attention graph pooling. In *International conference on machine learning* (pp. 3734–3743). PMLR.
Lešnik, S., Škrlj, B., Eržen, N., Bren, U., Gobec, S., Konc, J., et al. (2017). BoBER: Web interface to the base of bioisosterically exchangeable replacements. *Journal of Cheminformatics*, *9*(1), 1–8.
Li, X., & Cheng, Y. (2021). Understanding the message passing in graph neural networks via power iteration clustering. *Neural Networks*, *140*, 130–135.
Li, X., Yang, X., Wei, K., Deng, C., & Yang, M. (2022). Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9326–9335).
Li, S., Zhou, J., Xu, T., Dou, D., & Xiong, H. (2022). Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI conference on artificial intelligence (vol. 36), no. 4* (pp. 4541–4549).
Liu, Z., Fang, Y., Liu, C., & Hoi, S. C. (2021). Relative and absolute location embedding for few-shot node classification on graph. In *Proceedings of the AAAI conference on artificial intelligence (vol. 35), no. 5* (pp. 4267–4275).
Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S. J., et al. (2018). Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002.
Liu, S., Qu, M., Zhang, Z., Cai, H., & Tang, J. (2022). Structured multi-task learning for molecular property prediction. In *International conference on artificial intelligence and statistics* (pp. 8906–8920). PMLR.
Lu, B., Gan, X., Yang, L., Zhang, W., Fu, L., & Wang, X. (2022). Geometer: Graph few-shot class-incremental learning via prototype representation. arXiv preprint arXiv:2205.13954.
Lu, C., Liu, Q., Wang, C., Huang, Z., Lin, P., & He, L. (2019). Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI conference on artificial intelligence (vol. 33), no. 01* (pp. 1052–1060).
Luo, X., Ju, W., Qu, M., Chen, C., Deng, M., Hua, X.-S., et al. (2022). Dualgraph: Improving semi-supervised graph classification via dual contrastive learning. In *2022 IEEE 38th international conference on data engineering* (pp. 699–712). IEEE.
Luo, X., Ju, W., Qu, M., Gu, Y., Chen, C., Deng, M., et al. (2022). CLEAR: Cluster-enhanced contrast for self-supervised graph representation learning. *IEEE Transactions on Neural Networks and Learning Systems*.
Meanwell, N. A. (2011). Synopsis of some recent tactical application of bioisosteres in drug design. *Journal of Medicinal Chemistry*, *54*(8), 2529–2591.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.
National Center for Advancing Translational Sciences (2017). Tox21 challenge. http://tripod.nih.gov/tox21/challenge/. (Accessed 06 November 2016).
Nichol, A., Achiam, J., & Schulman, J. (2018). On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999.
Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., et al. (2020). Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1150–1160).
Rassil, A., Chougrad, H., & Zouaki, H. (2022). Augmented Graph Neural Network with hierarchical global-based residual connections. *Neural Networks*, *150*, 149–166.
Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., et al. (2016). ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chemical Research in Toxicology*, *29*(8), 1225–1251.
Rohrer, S. G., & Baumann, K. (2009). Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of Chemical Information and Modeling*, *49*(2), 169–184.
Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.
Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., & Borgwardt, K. M. (2011). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, *12*(9).
Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., & Borgwardt, K. (2009). Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics* (pp. 488–495). PMLR.
Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, *30*.
Sun, F.-Y., Hoffmann, J., Verma, V., & Tang, J. (2020). Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*.
Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1199–1208).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.
Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, *29*.
Von Oswald, J., Zhao, D., Kobayashi, S., Schug, S., Caccia, M., Zucchet, N., et al. (2021). Learning where to learn: Gradient sparsity in meta and continual learning. *Advances in Neural Information Processing Systems*, *34*, 5250–5263.
Wang, Y., Abuduweili, A., Yao, Q., & Dou, D. (2021). Property-aware relation networks for few-shot molecular property prediction. *Advances in Neural Information Processing Systems*, *34*, 17441–17454.
Wang, W., Guo, Y., Ju, W., Luo, X., & Deng, M. (2021). An interpretation of convolutional neural networks for motif finding from the view of probability. In *2021 IEEE 33rd international conference on tools with artificial intelligence* (pp. 176–183). IEEE.
Wang, S., Guo, Y., Wang, Y., Sun, H., & Huang, J. (2019). SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics* (pp. 429–436).
Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, *9*(2), 513–530.
Xie, Y., Zhang, Y., Gong, M., Tang, Z., & Han, C. (2020). Mgat: Multi-view graph attention networks. *Neural Networks*, *132*, 180–189.
Xie, G.-S., Zhang, Z., Xiong, H., Shao, L., & Li, X. (2022). Towards zero-shot learning: A brief review and an attention-based embedding network. *IEEE Transactions on Circuits and Systems for Video Technology*.
Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? In *ICLR*.
Xu, Z., Wang, S., Zhu, F., & Huang, J. (2017). Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 285–294).
Ye, H., Wang, Y., & Cao, F. (2021). A novel meta-learning framework: Multi-features adaptive aggregation method with information enhancer. *Neural Networks*, *144*, 755–765.
Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., & Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. *Advances in Neural Information Processing Systems*, *31*.
You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, *33*, 5812–5823.
Zablocki, E., Bordes, P., Soulier, L., Piwowarski, B., & Gallinari, P. (2019). Context-aware zero-shot learning for object recognition. In *International conference on machine learning* (pp. 7292–7303). PMLR.
Zhang, H., Li, H., & Koniusz, P. (2022). Multi-level second-order few-shot learning. *IEEE Transactions on Multimedia*.
Zhang, Z., Liu, Q., Wang, H., Lu, C., & Lee, C.-K. (2021). Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, *34*, 15870–15882.
Zhang, X., Wang, S., Zhu, F., Xu, Z., Wang, Y., & Huang, J. (2018). Seq3seq fingerprint: Towards end-to-end semi-supervised deep drug discovery. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 404–413).
Zhao, K., Zhang, Z., Jiang, B., & Tang, J. (2022). LGLNN: Label guided graph learning-neural network for few-shot learning. *Neural Networks*, *155*, 50–57.
Zheng, S., Yan, X., Yang, Y., & Xu, J. (2019). Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *Journal of Chemical Information and Modeling*, *59*(2), 914–923.