Robust Heterogeneous Graph Classification for Molecular Property Prediction with Information Bottleneck

Zhibin Ni[1], **Chang Liu**[1], **Hai Wan**[1], **Xibin Zhao**[1*]

[1] BNRist, KLISS, and School of Software, Tsinghua University
{nzb22,liuc24}@mails.tsinghua.edu.cn, {wanhai,zxb}@tsinghua.edu.cn

Abstract

Heterogeneous Graph Neural Networks (HGNNs) have achieved state-of-the-art performance in classifying molecular graphs, capitalizing on their ability to capture rich semantics. However, HGNNs for molecule property prediction exhibit significant susceptibility to adversarial attacks—a challenge that prior research has entirely overlooked. To fill this gap, this paper introduces the first study focused on *robust* graph-level representation learning tailored for heterogeneous molecular graphs. To achieve this goal, we propose a comprehensive **R**obust **H**eterogeneous **G**raph **C**lassification (**RHGC**) framework grounded in the Information Bottleneck principle, which aims to identify the most informative and least noisy heterogeneous subgraphs to derive *robust, holistic representations*. This is specifically accomplished through a dedicated *Node Semantic Purifier*, which enhances ***node-level*** and ***semantic-level robustness*** by eliminating label-irrelevant interference using graph stochastic attention and the Hilbert-Schmidt Independence Criterion, along with a *Global Graph Disentanglement* method, which improves ***graph-level robustness*** by addressing information leak. Experiments on three molecular benchmarks demonstrate that RHGC enhances accuracy by an average of 5.06% under all three attack settings and meanwhile by 4.33% on clean data.

Introduction

Molecular property prediction (Xu and Picek 2022) plays a vital role in drug discovery and virtual screening by automatically finding candidates with the expected properties of numerous molecules. With advances in deep learning (Yang et al. 2022; Yu and Wang 2024b), molecular property prediction has witnessed significant progress in recent years. Since molecules can be modeled as heterogeneous graphs with various types of atom and chemical bonds, Heterogeneous Graph Neural Networks (HGNNs) have achieved remarkable performance in molecular property prediction due to their encouraging capability to exploit rich inherent semantics (Zhao et al. 2022; Shi et al. 2023). Despite their promising performance, HGNNs for molecular property prediction require simultaneously modeling global structures and heterogeneous semantics, inevitably making them prone to adversarial attacks. For example, SiO_2 and TiO_2 share similar homogeneous structures and are both insoluble in water; however, replacing silicon with titanium alters their electrical properties: SiO_2 is an insulator, whereas TiO_2 is a semiconductor. Such vulnerabilities of HGNNs for molecular property prediction, however, have not yet been investigated in previous research.

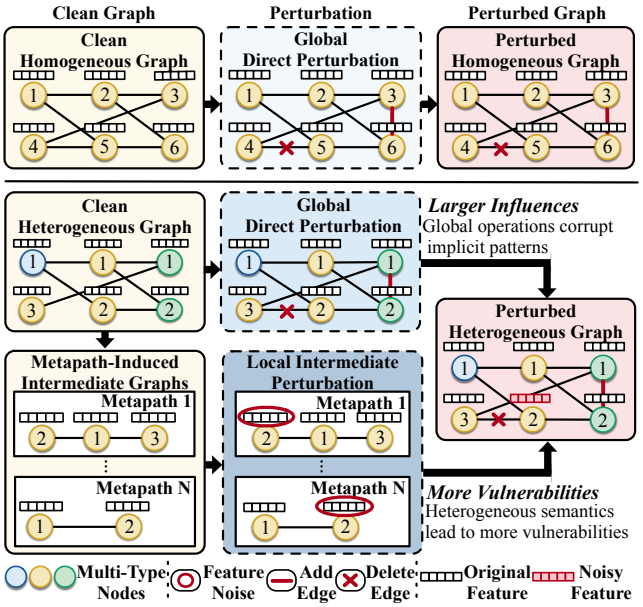

Figure 1: Comparison of adversarial attacks on homogeneous and heterogeneous GNNs. Heterogeneous GNNs experience larger global direct perturbation due to the simultaneous corruption of implicit pattern relying on heterogeneity. Besides, the heterogeneity introduces additional local intermediate perturbation, as it reveals more vulnerabilities that adversaries can exploit.

In particular, HGNNs for molecular property prediction face two robustness issues: *global direct perturbation with larger influences* and *local intermediate perturbation with more vulnerabilities*, as demonstrated in Fig. 1:

- Global direct perturbation (*i.e.,* the blue rectangles in Fig. 1) applied to heterogeneous molecular graphs has larger influences. Unlike homogeneous and other hetero-

*Corresponding author.

geneous graphs, heterogeneous molecular graphs contain implicit patterns dependent on heterogeneity, making them vulnerable to global direct perturbations. For example, bonding the carbonyl carbon in an aldehyde to another carbon instead of hydrogen forms a ketone group, significantly altering molecular properties;
- Local intermediate perturbation (*i.e.,* the dark blue rectangle in Fig. 1) introduced by heterogeneity exposes more vulnerabilities. The most popular variant of HGNNs, known as metapath-based HGNNs, adopts a hierarchical aggregation method (*i.e.*, including node-level and semantic-level) which transforms a heterogeneous graph into multiple intermediate graphs via metapaths. An adversary can easily devise strategies that target individual intermediate graphs when seeking optimal perturbations (Shang et al. 2023), thereby corrupting the semantics of molecular graphs.

Despite these critical robustness issues, to the best of our knowledge, there is no robust solution for heterogeneous graph classification. The most closely related methods are robust homogeneous graph classification methods (Sun et al. 2022; Ma et al. 2023; Seo, Kim, and Park 2024) and robust heterogeneous node classification methods (Yang et al. 2021; Zhang et al. 2022; Sang et al. 2023). However, directly applying them to heterogeneous graph classification is not favorable due to the following two challenges: on the one hand, previous works have overlooked constraining Y-irrelevant information at both node level and semantic level when learning robust heterogeneous node representations. In other words, existing methods follow the pipeline of metapath-based HGNNs, relying on the hypothesis that complementary information across metapaths is predictive and non-noisy. However, adversaries can inject noise to undermine such complementarity, thereby compromising the quality of node representations.

On the other hand, previous works often face information leak when optimizing robust structures at graph level. Specifically, in graph attack and defense, it is vital to mine the most robust structures for classification (Yu et al. 2021). The Information Bottleneck (IB) method offers a principled approach to this problem and has gained increasing attention. However, existing IB-based methods rely on trade-off hyper-parameters that fail to achieve exact information control, leading to Y-irrelevant data leak into final representations and thus degrading performance (Gabbay and Hoshen 2020; Pan et al. 2021a).

To address these two challenges, we propose a novel **R**obust **H**eterogeneous **G**raph **C**lassification (RHGC) model for molecular property prediction, which for the first time allows for yielding robust graph representations of heterogeneous graphs. This is achieved by leveraging the Information Bottleneck principle (Tishby, Pereira, and Bialek 1999), based on which we propose a novel hierarchical mechanism that disentangles Y-relevant information from Y-irrelevant information. The goal of this mechanism is to simultaneously provide robustness at the ***node-level***, ***semantic-level*** and ***graph-level***, covering a holistic design space of HGNNs.

To achieve this goal, RHGC comprises two novel modules. **First**, we introduce a dedicated *Node Semantic Purifier (NSP)* module to restrain Y-irrelevant information in both ***node-level*** and ***semantic-level*** aggregation. In particular, to learn robust heterogeneous structures, the NSP module utilizes graph stochastic attention at the node level to separate Y-relevant and Y-irrelevant subgraphs. Additionally, we employ the Hilbert-Schmidt Independence Criterion (HSIC) at the semantic level to enhance complementarity during the fusion of node representations derived from different metapaths, thereby yielding more robust node representations. **Second**, we further propose a *Global Graph Disentanglement (GGD)* module to optimize robust graph structures ***at graph level***. To prevent the leakage of Y-irrelevant information into the final graph-level representations, at the heart of GGD is a novel Disentangled Graph Information Bottleneck objective, which theoretically guarantees the achievement of accurate information control.

In sum, our contributions can be summarized as follows:
- To the best of our knowledge, we make the first attempt to reformulate molecular property prediction task as a *robust* heterogeneous graph classification problem, thereby extending the applicability of traditional methods to noisy scenarios.
- To address the *robust* heterogeneous graph classification problem, we propose RHGC, a holistic framework that for the first time simultaneously allows for node-level, semantic-level, and graph-level robustness in HGNNs, by employing a hierarchical mechanism grounded in the Information Bottleneck principle.
- Extensive experiments on three molecular datasets show that RHGC achieves an average improvement of 5.06% in classification accuracy across all three attack settings, and a 4.33% gain in performance on clean data.

Related Work

HGNN for Molecular Property Prediction. Molecular property prediction is a fundamental task for automatically screening target molecules with desirable properties. However, most existing molecular property prediction methods only consider neighboring atomic interactions in a homogeneous graph, ignoring different types of atomic nodes or edges. Recently, a new line of methods that use Heterogeneous Graph Neural Networks (HGNNs) to model the rich semantics in molecular graphs have achieved state-of-the-art performance (Long et al. 2021; Shi et al. 2023; Ji et al. 2023). Despite the superiority of HGNNs, they exhibit severe robustness weaknesses to adversarial perturbations in the molecular property prediction task. Unfortunately, these issues have been overlooked. To bridge this gap, we present a pilot study that, for the first time, makes it possible to learn robust graph-level representations tailored specifically for heterogeneous molecular graphs. Notably, we allow for holistic robustness, encompassing node-level, semantic-level, and graph-level aspects.

Robust Graph Classification. The robustness of graph classification models is crucial for ensuring highly reliable applications. Considering the vast importance, researchers have made many efforts to address the robustness challenge by employing data augmentation strategies (Wang

et al. 2021; Han et al. 2022), improving graph representation learning (You et al. 2020; Ma et al. 2023), advancing graph structure learning (Luo et al. 2021; Sun et al. 2022), and introducing graph Information Bottleneck principle (Wu et al. 2020; Seo, Kim, and Park 2024). In this paper, we make the the first attempt to propose robust heterogeneous graph classification. In addition, we propose a novel Disentangled Graph Information Bottleneck objective, which is able to avoid the information leak issue.

Preliminaries

Heterogeneous Graph and Metapath

A heterogeneous graph, denoted as $G = (\mathcal{V}, \mathcal{E})$, consists of a node set $\mathcal{V}$ and an edge set $\mathcal{E}$. A heterogeneous graph is also associated with a node type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{A}$ and an edge type mapping function $\psi : \mathcal{E} \rightarrow \mathcal{R}$. $\mathcal{A}$ and $\mathcal{R}$ denote the sets of predefined node types and edge types, where $|\mathcal{A}|+|\mathcal{R}| > 2$. For each edge type $R \in \mathcal{R}$, A_R denotes the corresponding binary adjacency matrix.

A metapath Φ is defined as a path which describes a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$, where $\circ$ denotes the composition operator on relations. For each metapath Φ_i, A_{Φ_i} denotes the metapath-induced binary adjacency matrix.

Problem Formulation

Given a set of heterogeneous molecular graphs $\mathcal{G} = \{(G_i)\}_{i=1}^N$, where $|\mathcal{G}| = N$, each G_i is assigned with a label $y_i \in \mathcal{Y} = \{1, 2, \cdots, C\}$ where C is the total number of classes. With paired heterogeneous graphs and labels $\{(G_i, y_i)\}_{i=1}^N$, the goal of robust heterogeneous graph classification is to learn a graph classifier $f : \mathcal{G} \rightarrow \mathcal{Y}$ which is able to obtain more robust representations and perform downstream graph-level classification task. During the test process, the classification performance of the learned graph classifier is expected to be robust against both targeted and non-targeted adversarial attacks.

The Proposed Model

In this section, we begin by providing an overview of our model, followed by the elaboration of the proposed individual NSP and GGD modules.

Model Overview

As shown in Fig. 2, the proposed RHGC consists of two modules: Node Semantic Purifier (NSP) and Global Graph Disentanglement (GGD). **First**, a node semantic purifier is proposed to generate robust heterogeneous structures and node representations. In the NSP module, graph stochastic attention is leveraged at the node level to disentangle Y-relevant from Y-irrelevant subgraphs, while HSIC is employed at the semantic level to optimize the complementarity of representations from different metapaths. **Second**, a global graph disentanglement method is proposed to alleviate the information leak at graph level. In the GGD module, the Y-relevant and Y-irrelevant subgraphs are optimized via novel Disentangled Graph Information Bottleneck. During training process, an adversarial training algorithm is proposed to encourage better disentanglement.

Node Semantic Purifier

Challenges Analysis. Prevalent metapath-based HGNNs learn the node representations on different metapath-induced intermediate graphs (node-level aggregation) and then use semantic fusion to learn the final node representations (semantic-level aggregation). Despite various robust solutions following this pipeline proposed, they *completely* neglect restraining Y-irrelevant information in both node-level and semantic-level aggregation. Different metapaths describe semantics in different views and the truly predictive information contained in different metapaths is complementary, which is called *complementary hypothesis* (Yang et al. 2021). In practice, however, adversaries can perturb metapath-induced intermediate graphs to accumulate noises at node-level and semantic-level aggregation, thereby corrupting the complementarity.

Method Rationale. To address the above challenge, we propose to concurrently achieve node-level and semantic-level robustness for blocking Y-irrelevant information. Towards this end, we develop a node semantic purifier (NSP) module based on the graph stochastic attention at node level to obtain robust graph structures, as well as Hilbert-Schmidt Independence Criterion (HSIC) at semantic level to further promote complementarity when aggregating representations from different metapaths.

Node-level Robustness. Specifically, at the node level, in a given metapath-induced adjacency matrix A_{Φ_i}, each node contains both Y-relevant and Y-irrelevant information, reflecting predictive signals and noise, respectively. We use two distinct heterogeneous graph convolution layers to separately obtain the Y-relevant and Y-irrelevant node representations $Z^T_{\Phi_i}$ and $Z^S_{\Phi_i}$. Furthermore, to generate Y-relevant and Y-irrelevant adjacency matrix from A_{Φ_i}, NSP learns two subgraph extractor g_{ϕ_t,Φ_i} and g_{ϕ_s,Φ_i} based on GSAT (Miao, Liu, and Li 2022). Taking subgraph extractor g_{ϕ_t,Φ_i} for instance, for each edge (u, v), the concatenated representations of the source node and the destination node $\left(z^T_{u,\Phi_i}, z^T_{v,\Phi_i}\right)$ are mapped to $p^T_{uv,\Phi_i} \in [0, 1]$. Finally, during the forward pass of each training iteration, each edge is sampled from a Bernoulli distribution $\alpha^T_{uv,\Phi_i} \sim \text{Bern}\left(p^T_{uv,\Phi_i}\right)$. To make the sampling process from p^T_{uv,Φ_i} differentiable, the Gumbel-Softmax reparameterization trick (Jang, Gu, and Poole 2016) is introduced. After this process, we can extract the Y-relevant adjacency matrix $A^T_{\Phi_i}$ by applying an attention-guided edge mask $\alpha^T_{\Phi_i}$ to original adjacency matrix, formulated as $A^T_{\Phi_i} = \alpha^T_{\Phi_i} \odot A_{\Phi_i}$. Here α^T_{uv,Φ_i} represents the importance of the edge (u, v) in the Φ_i-induced Y-relevant subgraph and $\odot$ is entry-wise product. In terms of Y-irrelevant subgraph extractor g_{ϕ_s,Φ_i}, we conduct the same process.

Consequently, by applying the aforementioned process in each metapath-induced graph, we are able to generate $A^T = \{A^T_{\Phi_{1:M}}\}$ and $A^S = \{A^S_{\Phi_{1:M}}\}$, which are represented in Fig. 2, respectively. This process aids in reducing noisy information propagation at the node level.

Semantic-level Robustness. At the semantic level, it is imperative to further enhance the complementarity of representations derived from different metapaths for the same node

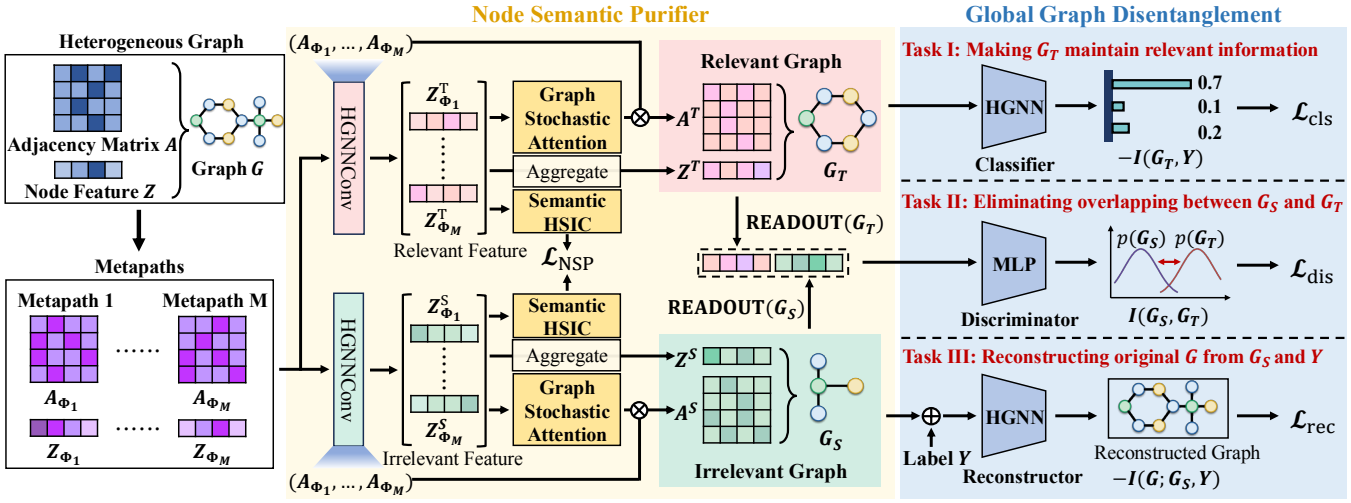

Figure 2: The main framework of the proposed RHGC for the molecular property prediction task. The framework consists of two modules, including a node semantic purifier module (the yellow part) and a global graph disentanglement module (the blue part). $I\left(G_T;Y\right)$, $I\left(G;G_S,Y\right)$, and $I\left(G_S;G_T\right)$ are different Mutual Information items defined in Eq. (4).

before semantic fusion. Given that HSIC has been both theoretically and empirically validated as an appropriate measure for assessing the (in)dependence between two signals, we employ HSIC as a regularization term to mine complementary information across different metapaths. Through the application of HSIC, we modulate the interdependencies among metapaths, thereby facilitating the adaptive extraction and fusion of information that significantly improves predictive performance. The HSIC formula is as follows:

$$HSIC(X,Y) = \frac{1}{(N-1)^2} tr(K_X H K_Y H), \quad (1)$$

where $H = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathbf{T}} (\in \mathbb{R}^{N\times N})$, with $\mathbf{I} (\in \mathbb{R}^{N\times N})$ denoting the identity matrix, and $\mathbf{1}$ representing a column vector with all elements being 1. K_X is the kernel matrix with elements $K_X(X_i, X_j) = \exp(\frac{||X_i, X_j||_2^2}{2\sigma^2})$, where σ is the bandwidth. We use the HSIC to formulate the semantic HSIC objective of Node Semantic Purifier as follows:

$$\begin{aligned}\mathcal{L}_{\text{NSP}} = &\sum_{i\neq j} HSIC(Z^T_{\Phi_i}, Z^T_{\Phi_j}) \\ &+ \sum_{i\neq j} HSIC(Z^S_{\Phi_i}, Z^S_{\Phi_j}),\end{aligned} \quad (2)$$

thereby eliminating noisy information at semantic level. Consistent with prior work (Wang et al. 2019), we employ semantic-level attention to derive the final Y-relevant and Y-irrelevant node representations, denoted as Z^T and Z^S, respectively.

Output of NSP. In conclusion, the robust node representation Z^T(*resp.*, Z^S) and Y-relevant (*resp.*, Y-irrelevant) adjacency matrix A^T(*resp.*, A^S) make up the Y-relevant (*resp.*, Y-irrelevant) subgraph G_T(*resp.*, G_S).

Global Graph Disentanglement

Challenges Analysis. After extracting the robust node representations Z^T, Z^S and the corresponding adjacency matrixes A^T, A^S by the NSP module, the framework still faces a great challenge of information leak when learning robust structures for graph classification, which lies in the fact that existing works inevitably lack precise information control at graph level. For example, IB-based works, which have seen a surge in recent years, identify an informative yet compressed subgraph G_T from the original graph G by optimizing the following objective:

$$\min_{G_T\in\mathbb{G}_{sub}} -I(Y;G_T) + \beta I(G;G_T), \quad (3)$$

where $\mathbb{G}_{sub}$ indicates the set of all subgraphs of G, Y is the label of G, $I(Y;G_T)$ and $I(G;G_T)$ denote the mutual information between Y and G_T, G and G_T, respectively, and β is the Lagrange multiplier that controls the trade-off between these two terms, which requires multiple attempts to find the optimal value in the conventional IB optimization process.

Method Rationale. By extending the disentangled information bottleneck (Pan et al. 2021b) to graph, we formally define the Disentangled Graph Information Bottleneck objective of GGD module as follows:

$$\mathcal{L}_{\text{GGD}} = -I\left(G_T;Y\right) - I\left(G;G_S,Y\right) + I\left(G_S;G_T\right). \quad (4)$$

The above objective can be divided into three tasks, we first maximize $I\left(G_T;Y\right)$ to ensure that Y can be accurately decoded from G_T, guaranteeing that G_T at least captures the information related to Y **(Task I)**. Next, we maximize $I\left(G;G_S,Y\right)$ to ensure that (G_S,Y) represents the overall information of G, thereby ensuring that G_S covers the Y-irrelevant information **(Task II)**. Finally, we minimize $I\left(G_S;G_T\right)$ to enforce the disentanglement between G_S and G_T, effectively separating Y-relevant information from Y-irrelevant information, thus tightening the information boundaries of the two subgraphs **(Task III)**. It can be seen that modeling the Y-irrelevant part into the IB objective simultaneously addresses the issues of information leak and multiple optimizations of β.

Optimization of Task I&II. To optimize the first two terms in Eq. (4), similar to previous works (Chalk, Marre, and Tkacik 2016; Achille and Soatto 2018; Kolchinsky, Tracey, and Wolpert 2019), we derive the variational lower bound for $I(G_T;Y)$ and $I(G;G_S,Y)$ terms as:

$$I(G_T;Y) \geq \mathbb{E}_{p(G_T,y)}\left[\log c(y|G_T)\right], \tag{5}$$

$$I(G;G_S,Y) \geq \mathbb{E}_{p(G,G_S,y)}\left[\log r(G|G_S,y)\right], \tag{6}$$

where $c(y|G_T)$ and $r(G|G_S,y)$ are two variational probabilistic mappings. The former can be easily modeled as a cross-entropy classification loss $\mathcal{L}_{clf}$ through G_T as:

$$\begin{aligned}\mathbb{E}_{p(G_T,y)}\left[\log c(y|G_T)\right] &= \mathcal{L}_{clf}\left(f\left(g_{\phi_t}(G)\right),y\right)\\ &= -\log f\left(g_{\phi_t}(G)\right)_y,\end{aligned} \tag{7}$$

where the variational probabilistic mapping $c(y|G_T)$ is parameterized by a graph classifier f, and g_{ϕ_t} represents subgraph extractors $g_{\phi_t,\Phi_{1:M}}$. While the latter is modeled as a graph reconstruction loss $\mathcal{L}_{rec}$ with G_S and label y as:

$$\begin{aligned}\mathbb{E}_{p(G,G_S,y)}\left[\log r(G|G_S,y)\right] &= \mathcal{L}_{rec}\left(R\left(g_{\phi_s}(G),y\right),G\right)\\ &= \frac{1}{\sum_{i=1}^{M}|\mathcal{V}_{\Phi_i}^S|^2}\sum_{i=1}^{M}\sum_{u,v}\left(A_{uv,\Phi_i}^S - \hat{A}_{uv,\Phi_i}^S\right)^2,\end{aligned} \tag{8}$$

where the variational probabilistic mapping $r(G|G_S,y)$ is parameterized by a graph reconstructor R, A and $\hat{A}$ representing adjacency matrix of the original and reconstructed graph.

Optimization of Task III. For the third term in Eq. (4), directly optimizing the objective $I(G_S;G_T) = D_{\mathrm{KL}}\left[p(G_S,G_T) \parallel p(G_S)\,p(G_T)\right]$ is intractable because both $p(G_S,G_T)$ and $p(G_S)\,p(G_T)$ contain many related random variables, making it difficult to eliminate all of them. Inspired by Generative Adversarial Networks (Goodfellow et al. 2014), GGD mitigates this by using adversarial training to reduce the gap between $p(G_S,G_T)$ and $p(G_S)\,p(G_T)$, thereby minimizing $I(G_S;G_T)$.

Specifically, GGD samples from the joint distribution $p(G_S,G_T)$ by uniformly sampling x from the dataset at random, then from the conditional distribution $p(G_S,G_T|x)$ (Kim and Mnih 2018). To approximate sampling from product of marginal distributions $p(G_S)\,p(G_T)$, we randomly shuffle the samples of $p(G_S,G_T)$ along the batch axis (Belghazi et al. 2018). Finally, GGD uses the density-ratio trick (Kim and Mnih 2018), introducing a discriminator d that estimates the probability that its input comes from $p(G_S,G_T)$ rather than $p(G_S)\,p(G_T)$ as:

$$\begin{aligned}\mathcal{L}_{dis}(G_S,G_T) = &\mathbb{E}_{p(G_S)p(G_T)}\log d(G_S,G_T)\\ &+\mathbb{E}_{p(G_S,G_T)}\log\left(1-d(G_S,G_T)\right).\end{aligned} \tag{9}$$

Adversarial training is performed during training process.

Loss Function

In our method, the loss $\mathcal{L}_{\mathrm{NSP}}$ from the NSP module and the Disentangled Graph Information Bottleneck loss $\mathcal{L}_{\mathrm{GGD}}$ from the GGD module are jointly optimized in adversarial training. The overall loss function is defined as:

$$\begin{aligned}\mathcal{L} &= \lambda\mathcal{L}_{\mathrm{NSP}} + \mathcal{L}_{\mathrm{GGD}}\\ &= \lambda\mathcal{L}_{\mathrm{NSP}} + \mathcal{L}_{clf} + \mathcal{L}_{rec} + \mathcal{L}_{dis}.\end{aligned} \tag{10}$$

Experiment Setup

In this section, we introduce the detailed experimental setup, including the datasets, baselines, metric, adversarial attack settings, and implementation details.

Datasets

Three datasets are used to evaluate the performance of all models, including Open Graph Benchmark (Hu et al. 2020) (OGB)-molbace, MUTAG (Rupp et al. 2012) and PROTEINS (Borgwardt et al. 2005). These are datasets related to molecules or bioinformatics, and are widely used for evaluations on graph classification.

Baselines and Metric

To verify the effectiveness of the proposed method, we select baselines from two related tasks: heterogeneous graph classification methods and robust homogeneous graph classification methods. For the heterogeneous graph classification methods, we select two baselines including muxGNN (Melton and Krishnan 2023) and HeGCL (Shi et al. 2023). For the robust homogeneous graph classification methods, we select two data augmentation methods including NodeSam (Yoo, Shim, and Kang 2022) and G-Mixup (Han et al. 2022), two graph representation learning methods including InfoGraph (Sun et al. 2019) and MGRL (Ma et al. 2023), and three graph information bottleneck methods including GIB (Wu et al. 2020), VIB-GSL (Sun et al. 2022) and PGIB (Seo, Kim, and Park 2024). For performance evaluation, we report accuracy for all datasets (Yu and Wang 2024a).

Adversarial Attack Settings

We compare baselines and the proposed RHGC under two adversarial attack settings.

- **Non-targeted Adversarial Attack:** We produce synthetic datasets by attacking graph structures and node features, respectively. **(1) Attack graph structures.** Following prior works (Tong et al. 2012), we choose nodes with top degrees and randomly insert/delete one edge among them. We select 20% of nodes in one graph to perturb. **(2) Attack node features.** We add random Gaussian noise $m \cdot r \cdot \epsilon$ to each dimension of the node features for all nodes, where r is the reference amplitude of original features, and $\epsilon \sim N(\mathbf{0},\mathbf{I})$. m acts as the parameter to control the feature adjusted magnitude. We set the value of m to 2.0 in the experiments.

Dataset	PROTEINS			OGB-molbace			MUTAG		
Model	**Clean**	**Structure Attack**	**Feature Attack**	**Clean**	**Structure Attack**	**Feature Attack**	**Clean**	**Structure Attack**	**Feature Attack**
HeGCL	77.37±1.53	68.12±4.90	74.12±3.01	83.61±4.32	71.09±3.98	74.01±3.68	86.36±1.82	76.92±3.71	79.53±4.12
muxGNN	77.41±2.42	70.21±3.14	71.54±4.23	82.15±3.01	73.56±1.04	74.93±2.13	86.53±1.97	77.21±2.63	78.21±3.21
NodeSam	63.78±1.04	59.21±0.94	60.58±1.63	62.75±2.16	57.21±1.98	58.21±1.37	75.67±1.34	64.25±1.20	65.78±2.13
G-Mixup	64.71±1.95	60.12±2.10	61.79±1.83	65.75±2.10	60.35±1.76	61.42±0.96	76.32±1.72	66.34±2.93	65.12±1.01
InfoGraph	63.98±1.18	58.45±2.78	59.51±1.65	64.75±2.15	59.77±3.26	60.81±1.22	76.09±1.32	66.13±2.96	69.13±1.69
MGRL	65.85±1.33	59.18±2.16	60.21±1.57	67.75±2.16	62.69±1.34	63.52±0.90	77.56±0.47	69.12±1.60	70.24±0.97
GIB	75.25±2.92	71.32±3.31	72.01±2.13	64.39±1.20	60.58±1.12	61.26±0.91	79.00±6.24	72.53±4.31	73.90±3.01
VIB-GSL	73.66±3.32	68.47±2.02	69.08±2.47	82.15±3.37	77.24±2.16	78.53±1.43	81.12±1.42	75.25±2.56	76.23±2.01
PGIB	77.50±2.38	72.19±1.47	73.06±2.02	83.45±3.47	76.32±2.19	78.82±1.26	85.50±5.22	79.13±3.12	79.25±2.67
RHGC	**81.24±2.11**	**77.15±2.62**	**77.63±2.36**	**87.87±1.65**	**82.03±2.19**	**83.41±1.01**	**91.52±2.23**	**85.63±3.62**	**86.10±2.13**

Table 1: Accuracy (% ± standard deviation) of graph classification task on real-world datasets against non-targeted adversarial attacks. The best results are shown in **bold type** and the runner-ups are underlined.

- **Targeted Adversarial Attack:** We focus on evasion attack, a typical type of targeted adversarial attack that perturbs graph in the test phase and guides the model to misclassify the target graph. We generate perturbations based on GRABNEL (Wan et al. 2021) in experiments as it can be easily adapted to perform various attacks on different metapaths under the black-box evasion attack setting. We set the attack budget of Δ (*i.e.* we are allowed to flip up to Δ edges from G) as $10\%|\mathcal{E}_i|$ for each graph G_i, where $|\mathcal{E}_i|$ denotes the number of edges in graph G_i.

Implementation Details

The proposed method is implemented with *PyTorch* 2.1 framework on Ubuntu 22.04. We set the maximum epoch number as 200 with the early stopping strategy (Xi et al. 2024). Each dataset is split into training, validation, and test sets with a ratio of 80%, 10%, and 10%, respectively. We use Adam (Fan et al. 2024a) with a learning rate selected from $\{1e$-03, $3e$-03, $1e$-04, $3e$-04$\}$ and adopt the grid search for the best performance using the validation split (Fan et al. 2024b). To mitigate the effects of random noise, we report the results from 10 runs with different random seeds.

Results and Analysis

In this section, we conduct experiments regarding robustness against both non-targeted and targeted adversarial attacks, ablation study, and hyper-parameter analysis to validate the proposed RHGC.

Against Non-targeted Adversarial Attacks

In this section, we evaluate model performance on the heterogeneous graph classification task, as well as the robustness against non-targeted adversarial attacks in terms of graph structures and node features. Results are concluded in Table 1. Here we have the following observations:

(1) The performance of all models dramatically drops under both feature and structure attacks, which demonstrates their common limitations. Adversarial attacks can significantly degrade the performance of heterogeneous graph classification baselines, with an average performance degradation of 8.12%, indicating that HGNNs experience larger influences from global direct perturbations. Although robust homogeneous graph classification methods exhibit some degree of resilience against adversarial attacks, they generally experience an accuracy decline of 5.68%. Furthermore, these methods often fall short in achieving precise information control and fail to address the additional vulnerabilities introduced by heterogeneity, resulting in suboptimal performance. On the contrary, RHGC consistently outperforms all other methods across all attack settings, with an average classification performance decline of just 4.94%. Simultaneously, RHGC exhibits an average improvement of 5.53% over the baselines under both structure and feature attacks. The results demonstrate RHGC's superior performance against both structure and feature attacks.

Dataset	Method	Clean	Evasion Attack
PROTEINS	muxGNN	77.41±2.42	52.98±3.69
	MGRL	65.85±1.33	53.78±2.01
	PGIB	77.50±2.38	58.14±4.49
	RHGC	**81.24±2.11**	**64.67±4.27**
OGB-molbace	muxGNN	82.15±3.01	57.74±4.20
	MGRL	67.75±2.16	52.08±3.32
	PGIB	83.45±3.47	63.53±4.57
	RHGC	**87.87±1.65**	**67.65±3.26**
MUTAG	muxGNN	86.53±1.97	65.35±4.53
	MGRL	77.56±0.47	64.22±3.12
	PGIB	85.50±5.22	70.29±3.68
	RHGC	**91.52±2.23**	**74.33±3.82**

Table 2: Results against evasion attack. The best results are shown in **bold type** and the runner-ups are underlined.

Method	PROTEINS	OGB-molbace	MUTAG
Baseline	69.64±4.71	77.21±3.46	80.88±8.94
NSP	74.72±3.54	80.23±4.21	84.33±3.71
GGD	78.12±1.96	81.61±2.76	86.49±2.43
NSP+GGD	**81.24±2.11**	**87.87±1.65**	**91.52±2.23**

Table 3: The ablation study results. The best results are shown in **bold type** and the runner-ups are underlined.

(2) It is significant that RHGC outperform all baselines on all clean datasets as well. Compared with the state-of-the-art baselines, the proposed method improves accuracy by 3.74%, 4.26%, and 4.99% on PROTEINS, OGB-molbace, and MUTAG, respectively. The results demonstrate the effectiveness of the proposed method for molecular property prediction. This is mainly due to the proposed NSP module and GGD module, which can effectively remove Y-irrelevant information and extract the robust graph representations with the proposed Disentangled Graph Information Bottleneck objective. Therefore, the proposed method can make full use of truly predictive information to enhance the performance of molecular property prediction.

Against Targeted Adversarial Attacks

In this section, we continue to compare RHGC with representative baselines standing out in the previous subsection, considering the classification performance and robustness against targeted adversarial attacks, which demonstrates whether RHGC successfully defends the attacks. Results are reported in Table 2.

Results reveal that all models experience a significant performance decline under evasion attacks, with baseline classification accuracy falling by an average of 18.40%, which is concerning as it suggests that seeking perturbations considering metapaths may result in more severe consequences. In contrast, RHGC shows an average decline of 17.99% and outperforms the baselines by 4.89% in accuracy under evasion attack mode. This shows the importance of enhancing robustness across all dimensions of the HGNN design space.

Ablation Study

In this section, to verify the effectiveness of each module in the proposed RHGC, ablation study is conducted on all datasets with different combinations of the key modules. Specifically, 4 combinations of key modules are compared in the ablation study as the followings:

- **Baseline:** The basic model of GSAT backbone.
- **NSP:** The baseline model with NSP module.
- **GGD:** The baseline model with GGD module.
- **NSP+GGD:** The proposed RHGC model.

As shown in Table 3, the GSAT backbone performs the worst, which is limited by the neglect of the heterogeneous nature and the insufficiency of modeling robust structures. Incorporating the GGD module into GSAT backbone significantly improves the classification performance, which indicates that the GGD module effectively disentangles the global Y-irrelevant information from the robust graph-level representation. Moreover, the classification performance is also improved through the incorporation of the NSP module, which demonstrates the effectiveness of constraining Y-irrelevant information during node-level and semantic-level aggregation. The proposed RHGC that combines NSP and GGD modules makes full use of the aforementioned advantages and achieves the best performance.

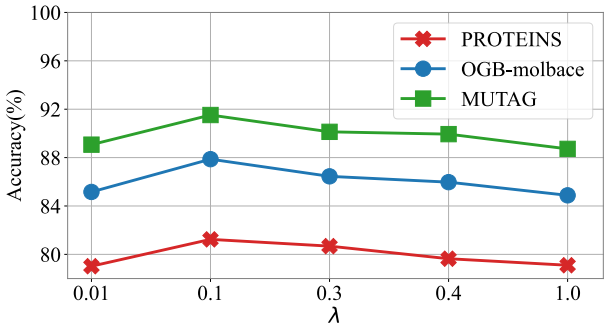

Figure 3: Hyper-parameter analysis results for different values of hyper-parameter λ on all datasets.

Hyper-parameter Analysis

In this section, we conduct a hyper-parameter analysis to evaluate the classification performance of the proposed RHGC method under varying values of the key hyper-parameter λ. The hyper-parameter analysis is performed across all datasets with λ set to $\{0.01, 0.1, 0.3, 0.4, 1\}$, as depicted in Fig. 3. It is observed that our method achieves the best performance when $\lambda = 0.1$ for all datasets. Therefore, we set the values of λ to 0.1 in the experiments.

Limitation Discussions

In this section, we discuss the limitations of our work. First, due to the limited research on poisoning attacks in graph classification and the unavailability of their code, we focus on evasion attacks in targeted adversarial scenarios. Additionally, to implement $I(G; G_S, Y)$, the reconstructor is adapted from the GAE (Kipf and Welling 2016). We leave the exploration of advanced graph reconstruction techniques for heterogeneous graphs as future work.

Conclusion

In this paper, we propose a novel RHGC method which is the first to enable *robust* heterogeneous graph representation learning for molecular property prediction. We introduce Node Semantic Purifier and Global Graph Disentanglement modules to effectively disentangle the Y-relevant and Y-irrelevant information at node level, semantic level and graph level, thereby covering the comprehensive design space of HGNNs. Extensive experiments on three molecular benchmarks demonstrate the superior robustness of RHGC over state-of-the-art baselines by a large margin in molecular property prediction tasks, both on noisy and clean data.

Acknowledgments

This research is sponsored in part by the NSFC Program (No. U20A6003), Industrial Technology Infrastructure Public Service Platform Project "Public Service Platform for Urban Rail Transit Equipment Signal System Testing and Safety Evaluation" (No. 2022-233-225), Science and technology innovation project of Hunan Province (No.2023RC4014).

References

Achille, A.; and Soatto, S. 2018. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2897–2905.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International conference on machine learning*, 531–540. PMLR.

Borgwardt, K. M.; Ong, C. S.; Schönauer, S.; Vishwanathan, S.; Smola, A. J.; and Kriegel, H.-P. 2005. Protein function prediction via graph kernels. *Bioinformatics*, 21: i47–i56.

Chalk, M.; Marre, O.; and Tkacik, G. 2016. Relevant sparse codes with variational information bottleneck. *Advances in Neural Information Processing Systems*, 29.

Fan, Z.; Wang, K.; Wen, K.; Zhu, Z.; Xu, D.; and Wang, Z. 2024a. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in neural information processing systems*.

Fan, Z.; Zhang, J.; Cong, W.; Wang, P.; Li, R.; Wen, K.; Zhou, S.; Kadambi, A.; Wang, Z.; Xu, D.; et al. 2024b. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems*.

Gabbay, A.; and Hoshen, Y. 2020. Demystifying Inter-Class Disentanglement. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, X.; Jiang, Z.; Liu, N.; and Hu, X. 2022. G-mixup: Graph data augmentation for graph classification. In *International Conference on Machine Learning*, 8230–8248. PMLR.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Ji, Y.; Wan, G.; Zhan, Y.; and Du, B. 2023. Metapath-fused heterogeneous graph network for molecular property prediction. *Information Sciences*, 629: 155–168.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International conference on machine learning*, 2649–2658. PMLR.

Kipf, T. N.; and Welling, M. 2016. Variational graph auto-encoders. *Bayesian Deep Learning Workshop (NIPS 2016)*.

Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2019. Nonlinear information bottleneck. *Entropy*, 21(12): 1181.

Long, Q.; Xu, L.; Fang, Z.; and Song, G. 2021. Hgk-gnn: Heterogeneous graph kernel based graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1129–1138.

Luo, D.; Cheng, W.; Yu, W.; Zong, B.; Ni, J.; Chen, H.; and Zhang, X. 2021. Learning to drop: Robust graph neural network via topological denoising. In *Proceedings of the 14th ACM international conference on web search and data mining*, 779–787.

Ma, G.; Hu, C.; Ge, L.; and Zhang, H. 2023. Multi-View Robust Graph Representation Learning for Graph Classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, 4037–4045.

Melton, J.; and Krishnan, S. 2023. muxGNN: Multiplex graph neural network for heterogeneous graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11067–11078.

Miao, S.; Liu, M.; and Li, P. 2022. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, 15524–15543. PMLR.

Pan, Z.; Niu, L.; Zhang, J.; and Zhang, L. 2021a. Disentangled information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9285–9293.

Pan, Z.; Niu, L.; Zhang, J.; and Zhang, L. 2021b. Disentangled information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9285–9293.

Rupp, M.; Tkatchenko, A.; Müller, K.-R.; and Von Lilienfeld, O. A. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5): 058301.

Sang, L.; Xu, M.; Qian, S.; and Wu, X. 2023. Adversarial heterogeneous graph neural network for robust recommendation. *IEEE Transactions on Computational Social Systems*, 10(5): 2660–2671.

Seo, S.; Kim, S.; and Park, C. 2024. Interpretable prototype-based graph information bottleneck. *Advances in Neural Information Processing Systems*, 36.

Shang, Y.; Zhang, Y.; Chen, J.; Jin, D.; and Li, Y. 2023. Transferable Structure-based Adversarial Attack of Heterogeneous Graph Neural Network. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2188–2197.

Shi, G.; Zhu, Y.; Liu, J. K.; and Li, X. 2023. Hegcl: Advance self-supervised learning in heterogeneous graph-level representation. *IEEE Transactions on Neural Networks and Learning Systems*.

Sun, F.-Y.; Hoffman, J.; Verma, V.; and Tang, J. 2019. InfoGraph: Unsupervised and Semi-supervised Graph-Level

Representation Learning via Mutual Information Maximization. In *International Conference on Learning Representations*.

Sun, Q.; Li, J.; Peng, H.; Wu, J.; Fu, X.; Ji, C.; and Philip, S. Y. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4165–4174.

Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The information bottleneck method. *Allerton*.

Tong, H.; Prakash, B. A.; Eliassi-Rad, T.; Faloutsos, M.; and Faloutsos, C. 2012. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 245–254.

Wan, X.; Kenlay, H.; Ru, R.; Blaas, A.; Osborne, M. A.; and Dong, X. 2021. Adversarial attacks on graph classifiers via bayesian optimisation. *Advances in Neural Information Processing Systems*, 34: 6983–6996.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032.

Wang, Y.; Wang, W.; Liang, Y.; Cai, Y.; and Hooi, B. 2021. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, 3663–3674.

Wu, T.; Ren, H.; Li, P.; and Leskovec, J. 2020. Graph information bottleneck. *Advances in neural information processing systems*, 33: 20437–20448.

Xi, H.; Nelson, J. D.; Hensher, D. A.; Hu, S.; Shao, X.; and Xie, C. 2024. Evaluating travel behavior resilience across urban and rural areas during the COVID-19 pandemic: contributions of vaccination and epidemiological indicators. *Transportation research part A: policy and practice*, 180: 103980.

Xu, J.; and Picek, S. 2022. Poster: clean-label backdoor attack on graph neural networks. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3491–3493.

Yang, L.; Wu, F.; Zheng, Z.; Niu, B.; Gu, J.; Wang, C.; Cao, X.; and Guo, Y. 2021. Heterogeneous Graph Information Bottleneck. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 1638–1645.

Yang, X.; Zhou, D.; Liu, S.; Ye, J.; and Wang, X. 2022. Deep model reassembly. *Advances in neural information processing systems*, 35: 25739–25753.

Yoo, J.; Shim, S.; and Kang, U. 2022. Model-agnostic augmentation for accurate graph classification. In *Proceedings of the ACM Web Conference 2022*, 1281–1291.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.

Yu, J.; Xu, T.; Rong, Y.; Bian, Y.; Huang, J.; and He, R. 2021. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*.

Yu, R.; and Wang, X. 2024a. Generator born from classifier. *Advances in neural information processing systems*, 36.

Yu, R.; and Wang, X. 2024b. Neural Lineage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4797–4807.

Zhang, M.; Wang, X.; Zhu, M.; Shi, C.; Zhang, Z.; and Zhou, J. 2022. Robust heterogeneous graph neural networks against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4363–4370.

Zhao, T.; Yang, C.; Li, Y.; Gan, Q.; Wang, Z.; Liang, F.; Zhao, H.; Shao, Y.; Wang, X.; and Shi, C. 2022. Space4hgnn: a novel, modularized and reproducible platform to evaluate heterogeneous graph neural network. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2776–2789.