Learning Topology-Specific Experts for Molecular Property Prediction

Suyeon Kim[1], Dongha Lee[2], SeongKu Kang[1], Seonghyeon Lee[1], Hwanjo Yu[1]*

[1]Pohang University of Science and Technology (POSTECH), South Korea
[2]University of Illinois at Urbana-Champaign (UIUC), United States
{kimsu, seongku, sh0416, hwanjoyu}@postech.ac.kr, donghal@illinois.edu

Abstract

Recently, graph neural networks (GNNs) have been successfully applied to predicting molecular properties, which is one of the most classical cheminformatics tasks with various applications. Despite their effectiveness, we empirically observe that training a single GNN model for diverse molecules with distinct structural patterns limits its prediction performance. In this paper, motivated by this observation, we propose TopExpert to leverage topology-specific prediction models (referred to as experts), each of which is responsible for each molecular group sharing similar topological semantics. That is, each expert learns topology-specific discriminative features while being trained with its corresponding topological group. To tackle the key challenge of grouping molecules by their topological patterns, we introduce a clustering-based gating module that assigns an input molecule into one of the clusters and further optimizes the gating module with two different types of self-supervision: topological semantics induced by GNNs and molecular scaffolds, respectively. Extensive experiments demonstrate that TopExpert has boosted the performance for molecular property prediction and also achieved better generalization for new molecules with unseen scaffolds than baselines. The code is available at https://github.com/kimsu55/ToxExpert.

Introduction

Molecular property prediction is one of the most classical cheminformatics tasks with applications in various areas such as quantum chemistry (Gilmer et al. 2017; Brockschmidt 2020) and biology (Jumper et al. 2021; Liu et al. 2020). It aims to clarify the correspondence between a molecule and its properties, including toxicity, barrier permeability, and adverse drug reactions. Molecular topology, which refers to the pattern of interconnections among atoms and thus determines the ultimate structure of a molecule (Rouvray 1986), has played an essential role in predicting the properties. It is well known that molecules with similar structural patterns are likely to have similar properties (Lo and Torres 2016), and accordingly, such topological information has been widely exploited for drug discovery and further modifications to improve target properties (Hu, Stumpfe, and Bajorath 2016; Naveja et al. 2019).
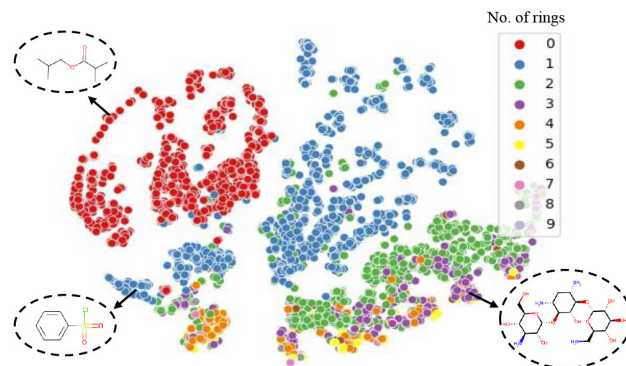
Figure 1: The t-SNE visualization of molecular representations produced by randomly-initialized GIN (Xu et al. 2019).

Recently, Graph neural networks (GNNs) have been successfully applied to predicting molecular properties (Yang et al. 2019, 2020; Yu and Gao 2022). Specifically, regarding molecules as graphs, GNNs extract localized spatial features by aggregating information at neighbor nodes and iteratively combining them to construct high-level representations, which effectively encodes the molecular topology (Figure 1). Several attempts have been made to help GNNs further capture topology information. One approach is to leverage frequently-occurred subgraphs, known as motifs, which are considered as basic meaning blocks with particular functions (Fey, Yuen, and Weichert 2020; Peng et al. 2020). Another approach is to preserve structural information in the process of aggregating node representations into a graph level (Baek, Kang, and Hwang 2021; Lee et al. 2021).

Despite their effectiveness, we observe that training a single GNN model to learn molecules with distinct structural patterns can hinder the model from capturing topology-specific features and eventually limits its performance. In Figure 2, for the property prediction of aromatic molecules, one GNN model trained on both aromatic and acyclic compounds (i.e., molecules having distinct topology) performs worse than another GNN model trained on aromatic compounds only despite its larger number of training data. In this sense, we argue that there is much room for improvement in case of utilizing topology-specialized models (called *ex-*
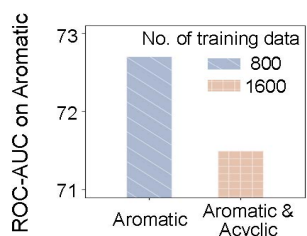
Figure 2: Classification performance on aromatic molecules.

perts) for each molecule group sharing similar topological semantics, instead of a single model.

One simple solution for grouping molecules by their topology is using *molecular scaffolds*. A scaffold is generally known as a core structure of a molecule, and each molecule is deterministically assigned to one of the scaffolds based on pre-defined rules. In particular, it has been well studied and exploited in pharmacological and biological areas (Hu, Stumpfe, and Bajorath 2016). However, there are two technical challenges to utilizing molecular scaffolds. First, scaffolds are too fine-grained to be directly used. For instance, there exist 7,831 molecules with 2,405 distinct scaffolds in the Tox21 dataset. Thus, assigning an expert to each scaffold is not feasible, as the number of molecules in each scaffold is too few to train an expert; only 3.3 molecules share the same scaffold on average for Tox21 dataset. Second, at the deployment of the model, it will face the challenge of handling new molecules of *unseen* scaffolds that are not in the training set. This arises from the scaffold splitting procedure, which enforces training and testing datasets sharing no common scaffolds; this has been widely used as a default evaluation setup because it provides a more realistic estimate of the model performance than random splitting (Wu et al. 2018). In this regard, simply using scaffold indices as the input features is not applicable because unseen scaffolds cannot be handled during testing.

In this paper, we propose a topology-specific expert model, named **TopExpert**, that exploits multiple experts specialized in their corresponding topological groups (Figure 3). Each expert is responsible for predicting the property of molecules in each group by effectively capturing discriminative features specific to its associated topological group; it serves as topology-specific knowledge. For the assignment of molecules to the experts, we introduce a novel gating module that identifies topological groups relevant to an input molecule based on its cluster assignment.

To obtain effective clusters of molecules based on their topology, we optimize our gating module in a self-supervised manner, in order to make molecular representations distinguishable according to their topological semantics. By using both topological features (encoded by a GNN) and prior knowledge (induced by scaffolds) of molecules, we strengthen (1) the cohesion of each cluster and (2) the alignment between fine-grained scaffolds and coarse-grained clusters. In the end, TopExpert clusters molecules having similar topological semantics into the same group, which allows us to identify relevant groups even for a new
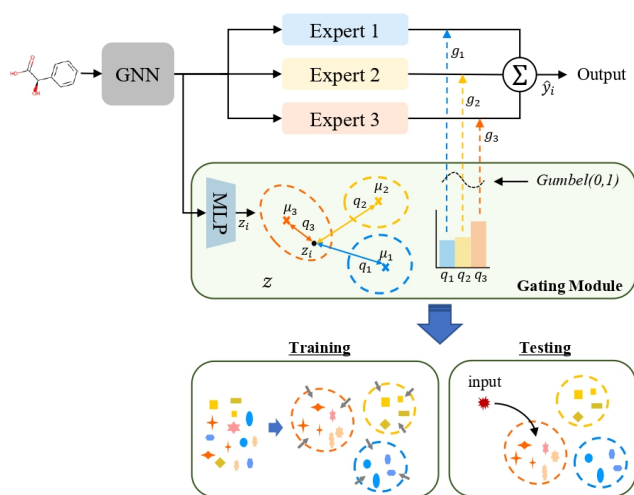

Figure 3: The overall framework of TopExpert. It consists of three components: GNN, a gating module, and multiple experts. In training, the gating module optimizes molecular representations to be distinguishable according to their topological information. In testing, it selects experts more relevant to an input molecule with a higher weight.

molecule with an unseen scaffold. The contribution of this work is threefold as follows:

- We demonstrate that learning molecules having distinct structural patterns with a single model may negatively affect capturing topology-specific features for property prediction, which has not been studied in previous works.
- We propose topology-specific experts with a novel clustering-based gating module that leverages topological information induced by GNNs and molecular scaffolds. To the best of our knowledge, this is the first work to incorporate molecular scaffolds into GNNs.
- We validate the effectiveness of TopExpert as a general tool for boosting the performance of existing GNN models through extensive experiments. Furthermore, we show that our model can be well-generalized to new molecules with unseen scaffolds.

Related Work

Molecular Property Prediction. It aims to clarify the correspondence between a molecule and its properties. For encoding molecules as numerical features, traditional approaches use hand-engineered molecular features such as molecular fingerprints (Rogers and Hahn 2010) and molecular descriptors (Hansen et al. 2015). However, such hand-crafted features are insufficient to reflect structural similarities and biological activities of molecules. Recently, GNNs have gained significant attention because of their convincing performance (Gilmer et al. 2017; Lu et al. 2019; Klicpera, Groß, and Günnemann 2020). GNNs also have the advantage that they can be directly applied to non-euclidean graph data without feature engineering. To better exploit molecular topology, several studies have leveraged motifs (i.e.,

frequently-occurred subgraphs) mostly obtained based on prior knowledge (Fey, Yuen, and Weichert 2020; Yu and Gao 2022). In addition, some attempts have been made to better preserve structural information in the process of aggregating node-level representations into a graph-level one (Baek, Kang, and Hwang 2021; Lee et al. 2021). Another line of work has focused on pre-training and self-supervised learning of GNNs to improve the performance on downstream tasks (Sun et al. 2019; Hu et al. 2020; Subramonian 2021).

Although the aforementioned methods have shown promising results with the help of their enhanced GNNs, they mainly utilize a single GNN model for their target task. However, our empirical findings imply that such a single model is insufficient to handle molecules with distinct structural patterns. For this reason, this work focuses on utilizing multiple experts to capture topological group-specific knowledge for property prediction, which is orthogonal to the existing studies.

Expert Models. It mainly employs a divide-and-conquer strategy for systems composed of many distinct experts, each of which learns to handle a part of input space (Jacobs et al. 1991). Recently, this idea of mixture of experts (MoE) has been combined with neural networks, and Shazeer et al. (2017) firstly propose a sparsely-gated MoE with the goal of reducing computational costs. In particular, MoE architecture has been validated in various applications such as image recognition (Ahmed, Baig, and Torresani 2016), machine translation (Shen et al. 2019), and recommender system (Kang et al. 2020, 2021). When it comes to graph classification, Hu et al. (2021) adopts a multi-gate MoE (Ma et al. 2018) to tackle the performance deterioration caused by class-imbalance. However, they do not consider the explicit graph topology in the gating process, and its performance is degraded when using a single gating network (Hu et al. 2021). Our work is distinguished from the above approaches in that we elaborate MoE to leverage topology knowledge induced by GNNs and molecular scaffolds in the gating process.

Preliminaries

Problem Formulation

In molecular property prediction, a molecule is usually regarded as a graph whose nodes and edges respectively correspond to the atoms and chemical bonds in the molecule. Given a training set $D = \{(G_i, y_i)\}_{i=1}^N$, where $G_i \in \mathcal{G}$ denotes a graph (i.e., molecule) and $y_i \in \{0, 1\}^T$ is its binary labels (i.e., target properties) for T tasks, the goal of our work is to train a model $f : \mathcal{G} \to \mathcal{Y}$ to predict properties of new input molecules. A graph G with n nodes and m edges is defined by its node features $X \in \mathbb{R}^{n \times d_{node}}$, edge features $E \in \mathbb{R}^{m \times d_{edge}}$, and the adjacency matrix $A \in \mathbb{R}^{n \times n}$, where $A(u, v) = 1$ if there is an edge between node u and node v, and $A(u, v) = 0$ otherwise.

Molecular Scaffolds

A scaffold is generally known as a core structure of a molecule (Hu, Stumpfe, and Bajorath 2016), and "Bemis and Murcko (BM) scaffolds" whose major building blocks are ring systems with linkers have been widely used (Bemis and Murcko 1996).[1] Note that the scaffold of a molecule is easily determined based on pre-defined rules. Scaffold-inducing structural information is instructive when exploiting a direct structure-property relationship in pharmacological and biological areas. For example, it was reported that 5,120 known drugs were categorized into 1,179 scaffolds, and 50 percent of the total drug molecules belong to only 32 of 1,179 (Bemis and Murcko 1996). In this sense, benchmark datasets for molecular property prediction are mainly split into training and testing sets to have disjoint scaffold indices, which makes the prediction more challenging yet realistic compared to random split (Wu et al. 2018).

Graph Neural Networks

To encode useful information from graph-structured data, GNNs iteratively compute node representations by aggregating messages from neighbor nodes and connected edges in a layer-wise manner (Scarselli et al. 2008). The representation of node u at the $(l+1)$-th GNN layer (denoted by $\mathrm{h}_u^{(l+1)}$) is defined by

$$\mathrm{h}_u^{(l+1)} = f_u^{(l)} \left(\mathrm{h}_u^{(l)}, f_a^{(l)} \left(\{(\mathrm{h}_v^{(l)}, \mathrm{h}_u^{(l)}, \mathrm{e}_{uv}) : \forall v \in \mathcal{N}(u)\} \right) \right), \tag{1}$$

where $\mathcal{N}(u)$ is the set of neighbors of node u, and $f_u^{(l)}$ and $f_a^{(l)}$ are the arbitrary differentiable functions for updating representations and aggregating messages, respectively. In the end, the graph-level representation h_G is obtained by summarizing all node representations using a readout function f_r such as averaging, summation, or sophisticated methods,

$$\mathrm{h}_G = f_r \left(\{\mathrm{h}_u^l | u \in G\} \right). \tag{2}$$

TopExpert: The Proposed Method

In this section, we present TopExpert, a GNN-based framework that adopts multiple experts for molecular property prediction. Each expert learns from molecules in the corresponding topological group, where the molecules are assigned by a clustering-based gating module.

Overview

The goal of TopExpert is to effectively capture topology-specific features by exploiting multiple experts specialized in their corresponding topological groups. In Figure 3, TopExpert consists of three parts: a GNN, K number of experts, and a gating module. Serving as a feature extractor, the GNN maps an input molecule into a low-dimensional molecule representation. Then, each expert, implemented as a fully-connected (FC) layer, computes its output (i.e., classification logits) by using the molecule representation. In the end, the final output is obtained by aggregating the experts' predictions with their weight values, which are calculated by the gating module to assign the input molecule into the topological groups (i.e., molecule clusters). To encourage the gating module to effectively assign each molecule into the clusters based on topological similarity, we adopt two learning

---

[1]In this paper, we refer to BM scaffolds as scaffolds for brevity.

objectives to make the molecule representations discriminative in terms of topological patterns: (1) The *clustering loss* strengthens the cohesion of the clusters by increasing the maximum cluster membership of each molecule, and (2) the *alignment loss* further aligns the coarse-grained clusters with the fine-grained scaffolds to encode topological semantics induced by molecular scaffolds.

Multiple Experts Learning for Property Prediction

The key idea of our gating module is to assign each molecule into the experts based on its similarity to the corresponding molecule clusters. To this end, given an input molecule G_i, the gating module non-linearly transforms its GNN-induced molecule representation h_{G_i} into the topology representation $z_i \in \mathcal{Z}$, which is used for molecule clustering based on topological patterns; we simply use a multi-layer perceptron (MLP) with two hidden layers: $z_i = \mathrm{MLP}(\mathrm{h}_{G_i}) \in \mathbb{R}^{d_{\mathcal{Z}}}$.

Then, to calculate the cluster assignment probability, we introduce trainable parameters for K cluster centroids, $M = [\mu_1; ...; \mu_K] \in \mathbb{R}^{K \times d_{\mathcal{Z}}}$.[2] To compute the assignment probability of the i-th molecule to the k-th topological cluster, denoted by q_{ik}, we measure the similarity between the topology representation z_i and the k-th centroid μ_k using a Student's t-distribution with one degree of freedom as a kernel (Van der Maaten and Hinton 2008):

$$q_{ik} = \frac{(1 + ||z_i - \mu_k||^2)^{-1}}{\sum_{k'} (1 + ||z_i - \mu_{k'}||^2)^{-1}}. \tag{3}$$

Based on the cluster assignment probability, we additionally adopt a stochastic selection strategy to help the experts to learn comprehensive topological patterns in the early training stage. Precisely, we use the Gumbel-Softmax, (Jang, Gu, and Poole 2017) which approximates the sampling from the cluster assignment distribution with the reparameterization trick for its differentiation:

$$g_{ik} = \frac{\exp\left((\log q_{ik} + \gamma_{ik})/\tau\right)}{\sum_{k'} \exp\left((\log q_{ik'} + \gamma_{ik'})/\tau\right)}, \tag{4}$$

where γ_{ik} is drawn from the standard Gumbel distribution; i.e., $\gamma_{ik} = -\log(-\log(U_{ik}))$ with $U_{ik} \sim Uniform(0, 1)$. The temperature of the softmax τ is initially set to the value T_0 and exponentially annealed down to the last value T_E; i.e., $\tau = T_0 (\frac{T_E}{T_0})^{\frac{e}{E}}$ such that $T_0 \gg T_E$ with the total training epochs E and the current epoch e. The higher τ value makes the Gumbel-Softmax distribution more uniform over the experts (Jang, Gu, and Poole 2017), which allows experts to stochastically explore diverse topological patterns in the early training phase (Balın, Abid, and Zou 2019).

Finally, each of the K experts outputs the logit $\mathrm{o}_{ik} \in \mathbb{R}^T$, and the final prediction output $\hat{y}_i \in \mathbb{R}^T$ is obtained by

$$\hat{y}_i = \sum_{k=1}^{K} g_{ik} \cdot \sigma(\mathrm{o}_{ik}), \tag{5}$$

---

[2]The centroids are initialized by K-means clustering on topology representations from a randomly-initialized model.

where $\sigma(\cdot)$ is a sigmoid function. To optimize the model parameters for the target prediction tasks (i.e., multi-label binary classification), we define the binary cross entropy loss:

$$\mathcal{L}_{class} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{BCE}(y_i, \hat{y}_i). \tag{6}$$

Note that, as training progresses, the weight distribution over all the experts, i.e., g_i, is getting closer to one-hot distribution due to the annealed temperature τ as in Eq. (5). In the end, each expert is only updated by gradients from its corresponding topology group, while gradients from the other groups hardly flow back to the expert with negligible weights. This encourages each expert to learn discriminative features from the corresponding subset of the data.

The remaining challenge is guiding the gating module to discover topological clusters of molecules, which should be effective for the target prediction tasks. Since there do not exist guidance labels for clustering, we optimize the topology representations and cluster centroids in a self-supervised manner by leveraging topological features (encoded by a GNN) and prior knowledge (induced by scaffolds).

Gating Module Learning for Cohesive Clustering

For the guidance of the gating module in terms of clustering, TopExpert directly optimizes the cluster assignment distribution q_i (Eq. (3)) with the help of the target distribution p_i computed using the current cluster assignment. As shown in Figure 1, GNNs are inherently capable of extracting discriminative structural features to some extent. Therefore, we define a target distribution for each molecule based on the cluster assignment of the current model, which therefore enhances the high-confident assignment (Xie, Girshick, and Farhadi 2016) via a squaring operation:

$$p_{ik} = \frac{q_{ik}^2 / \sum_i q_{ik}}{\sum_{k'} (q_{ik'}^2 / \sum_i q_{ik'})}. \tag{7}$$

Note that this target distribution strengthens the cluster cohesion, which tunes topology representations further discriminative with respect to their topological patterns.

To sum up, we define the clustering loss that enhances the cohesion of the clusters by minimizing KL divergence between a cluster assignment distribution q_i and a target distribution p_i as follows:

$$\mathcal{L}_{cluster} = \mathrm{KL}(P||Q) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \log \frac{p_{ik}}{q_{ik}}. \tag{8}$$

Gating Module Learning for Scaffold Alignment

To incorporate rich topological information into molecule clustering, we take advantage of molecular scaffolds as prior knowledge to guide the clustering process. Since scaffolds are much finer-grained than molecular clusters of our gating module, we propose a scaffold alignment scheme that aligns each molecule's scaffold index with its cluster assignment distribution. This facilitates the densely-located topology representations of molecules having not only the same scaffolds but also topologically similar scaffolds.

Inspired by the optimal transport (OT) problem (Monge 1781), we devise a distance measure between the cluster assignment distribution q_i and the scaffold distribution s_i for a molecule G_i. Note that OT has been widely used to measure the distance between two probability measures by optimizing all possible probabilistic couplings between them (Ma and Chen 2021; Lee et al. 2022). In the case of a training set with V scaffolds, the scaffold index of each molecule $s \in \{0,1\}^V$ is represented as a one-hot distribution over all the scaffolds. To formulate our OT problem, we define the cost matrix $M \in \mathbb{R}^{V \times K}$ between V scaffolds and K clusters based on the cosine distance of their representations; that is, $m_{vk} = 1 - \cos(e_v, \mu_k)$, where $e_v \in \mathbb{R}^{d_z}$ is the scaffold representation with trainable parameters. Using the cost matrix M with two distributions s and q, our distance is defined by

$$d_M(s, q) := \min_{T \in U(s,q)} \langle T, M \rangle. \qquad (9)$$

$\langle \cdot, \cdot \rangle$ is the Frobenius dot-product and $T \in \mathbb{R}^{V \times K}$ is one of the feasible transport matrices in the space $U(s, q) := \{T \in (\mathbb{R}^+)^{V \times K} | T\mathbf{1}_K = s, T^\top \mathbf{1}_V = q\}$, where $\mathbf{1}_d$ is the d-dimensional one vector. Since a scaffold distribution s is a one-hot distribution, Eq. (9) has the closed-form solution for the transport matrix without requiring any optimization process such as Sinkhorn algorithm (Cuturi 2013); i.e., $T^* = sq^\top$. In the end, we define the alignment loss to minimize the distance $d_M(s, q)$, which can be rewritten as follows:

$$\mathcal{L}_{align} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} s_{iv} \cdot q_{ik} \cdot (1 - \cos(e_v, \mu_k)). \quad (10)$$

Minimizing the distance for a molecule G_i has the effect of matching the molecule's topology representation z_i with its corresponding scaffold representation e_v while both representations approach the centroid of a specific cluster. When it comes to the entire training molecules, it aligns each scaffold with one of the clusters by gathering molecules with the same scaffold into the same cluster that best represents the topological semantics of the scaffold. In addition, the condition $V \gg K$ induces topologically similar scaffolds to be grouped together.

Remarks. We also considered two alternative strategies to utilize molecular scaffolds: one is minimizing the distance between a molecule's topology representation and its corresponding scaffold representation directly, e.g., $\sum_v s_{iv} \cdot ||z_i - e_v||_2$, and the other is applying scaffold classification loss by using the scaffolds as the label. We found that the OT-based alignment strategy consistently obtains higher performance than the alternatives (This will be discussed in Appendix B). Furthermore, the clustering part in TopExpert (i.e., GNN, MLP, Eq. (8), and Eq. (10)) can be used as a standalone model for deep molecule clustering. We name it *topology-based deep clustering*, and further details are provided in Appendix C.

Optimization and Inference

All the parameters in TopExpert, including the GNN, the experts, and the gating module, are optimized in an end-to-

end manner by minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{class} + \alpha \mathcal{L}_{cluster} + \beta \mathcal{L}_{align}, \qquad (11)$$

where α and β are the balance parameters that control the quality of the final topological clusters.

To make a deterministic prediction, a new input molecule at test time should be assigned into the clusters in a deterministic way. Thus, it computes the final prediction by using Eq. (5) that replaces g_{ik} with q_{ik}. In other words, the cluster assignment probability q_{ik} (in Eq. (3)) is directly utilized as the weight for the experts, without applying Gumbel-Softmax to get g_{ik} (in Eq. (4)).

Experiments

In this section, we focus on the following three research questions to validate the effectiveness of TopExpert.

- **RQ1** Does TopExpert effectively predict molecular properties by leveraging the underlying topological patterns?

- **RQ2** Does TopExpert provide robust predictions for unseen molecules by capturing topological patterns?

- **RQ3** Do $\mathcal{L}_{cluster}$ and $\mathcal{L}_{align}$ provide useful supervision with the gating module during training?

Experimental Settings

Dataset. We use 8 benchmark datasets (Wu et al. 2018) for molecular property prediction, and the statistics of the datasets are summarized in Table 1. Following the previous study (Hu et al. 2020), we extract molecular features of nodes, edges, and scaffold indices through RDKit.[3] For each dataset, we follow the scaffold splitting protocol (Hu et al. 2020), which sorts all the molecules by their scaffold indices and then splits them into training/validation/testing sets with a ratio of 80:10:10, respectively.

Baselines. We compare TopExpert with various baselines. All compared methods use GNN as a feature extractor, but they employ different strategies for making predictions.

- **SingleCLF** is the standard classifier that uses a classification module (i.e., a single expert).

The methods below employ multiple experts but use different gating schemes to consolidate the experts.

- **Mixture of Experts (MoE)** employs MLP with Gumbel-Softmax to select relevant experts for each molecule.

- **Expert-Ensemble (E-Ensemble)** (Dietterich 2000) uses the arithmetic mean to combine the outputs from experts.

- **GraphDIVE** (Hu et al. 2021) uses a weighted sum of the outputs from the experts. The weights are computed by a linear layer with Softmax.

To evaluate the effectiveness of TopExpert with a broad range of GNN backbones (Yang, Ma, and Cheng 2021), we use various GNN architectures: **GCN** (Kipf and Welling 2017), **GraphSAGE** (Hamilton, Ying, and Leskovec 2017), **GAT** (Veličković et al. 2017), **GIN** (Xu et al. 2019), and **Pre-trained GIN** (Hu et al. 2020). For all experiments, we employ 5-layer GNNs and each expert of a single FC layer.

---

[3]http://www.rdkit.org

GNN	Method	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	AVG.
	# of Mol.	2,039	7,831	8,575	1,427	1,478	93,087	41,127	1,513	-
	# of Tasks	1	12	617	27	2	17	1	1	-
GCN	SingleCLF	65.9 ± 0.9	74.4 ± 0.6	**63.6 ± 1.1**	60.6 ± 0.8	55.4 ± 3.6	74.0 ± 1.3	75.2 ± 1.4	71.0 ± 4.6	67.5
	MoE	65.4 ± 1.6	74.3 ± 0.4	61.5 ± 0.9	**60.8 ± 1.0**	68.1 ± 2.4	73.9 ± 1.2	75.5 ± 1.0	75.8 ± 2.8	69.4
	E-Ensemble	65.8 ± 2.8	74.4 ± 0.6	61.6 ± 0.7	60.3 ± 0.9	**70.5 ± 5.9**	74.1 ± 1.0	74.9 ± 1.0	76.0 ± 0.2	69.7
	GraphDIVE	62.5 ± 1.9	73.2 ± 0.9	62.0 ± 0.7	51.9 ± 2.0	60.4 ± 6.2	64.6 ± 4.7	65.2 ± 4.1	59.4 ± 4.6	62.4
	TopExpert	**67.0 ± 2.4**	**75.3 ± 0.4**	63.1 ± 0.6	**60.8 ± 0.9**	70.1 ± 6.0	**74.6 ± 0.6**	**76.6 ± 0.7**	**76.4 ± 1.9**	**70.5**
GraphSAGE	SingleCLF	**68.1 ± 1.5**	74.2 ± 0.8	**63.6 ± 0.7**	59.7 ± 1.0	53.4 ± 2.4	74.5 ± 2.5	74.6 ± 1.5	70.8 ± 3.3	67.4
	MoE	66.9 ± 2.0	74.5 ± 0.4	62.9 ± 0.7	61.7 ± 1.3	60.3 ± 4.4	73.0 ± 1.6	73.5 ± 1.0	71.1 ± 3.1	68.0
	E-Ensemble	67.3 ± 1.6	74.5 ± 0.5	62.4 ± 0.6	59.6 ± 0.8	**60.9 ± 3.2**	73.6 ± 6.1	75.0 ± 1.1	70.1 ± 2.5	67.9
	GraphDIVE	61.3 ± 2.3	**74.6 ± 0.5**	62.3 ± 0.6	57.1 ± 2.8	57.1 ± 7.6	68.2 ± 3.9	65.2 ± 2.8	65.9 ± 5.2	64.0
	TopExpert	67.6 ± 2.0	74.3 ± 0.5	62.6 ± 0.7	**62.6 ± 0.9**	58.7 ± 2.7	**76.0 ± 1.5**	**75.5 ± 1.0**	**74.0 ± 3.1**	**68.9**
GAT	SingleCLF	64.9 ± 1.2	**75.0 ± 0.8**	**63.5 ± 1.6**	61.0 ± 1.1	58.9 ± 1.4	**74.5 ± 0.9**	75.5 ± 1.7	75.3 ± 2.4	68.6
	MoE	64.0 ± 1.4	70.8 ± 0.7	62.3 ± 0.8	60.0 ± 1.6	54.1 ± 4.7	73.1 ± 1.8	73.0 ± 1.9	**76.5 ± 2.8**	66.7
	E-Ensemble	**66.8 ± 1.5**	72.2 ± 1.2	62.5 ± 0.6	59.4 ± 4.1	58.7 ± 4.4	73.5 ± 1.5	75.2 ± 1.1	77.0 ± 3.2	68.2
	GraphDIVE	64.1 ± 1.4	70.1 ± 1.3	60.4 ± 1.3	53.7 ± 1.7	**60.2 ± 7.2**	73.1 ± 1.6	75.5 ± 1.4	68.4 ± 7.5	65.7
	TopExpert	65.4 ± 2.1	74.9 ± 0.8	62.9 ± 0.9	**62.0 ± 1.3**	59.1 ± 2.5	74.1 ± 1.1	**77.3 ± 1.3**	76.3 ± 2.0	**69.0**
GIN	SingleCLF	68.9 ± 1.2	74.3 ± 0.6	**64.1 ± 1.6**	58.1 ± 1.5	58.8 ± 5.7	**76.1 ± 1.3**	75.6 ± 1.6	69.0 ± 4.7	68.1
	MoE	66.3 ± 2.0	74.5 ± 0.5	60.1 ± 1.0	58.6 ± 0.9	55.5 ± 3.0	**76.1 ± 0.8**	71.4 ± 2.7	68.8 ± 3.9	66.4
	E-Ensemble	66.5 ± 2.0	74.4 ± 1.1	60.7 ± 1.1	56.1 ± 1.6	59.8 ± 7.2	72.8 ± 2.5	76.2 ± 1.1	68.3 ± 5.2	66.9
	GraphDIVE	65.0 ± 2.6	72.1 ± 3.0	54.7 ± 1.2	52.9 ± 2.3	52.9 ± 6.9	65.5 ± 7.0	68.9 ± 2.1	62.5 ± 4.7	61.8
	TopExpert	**70.0 ± 0.7**	**75.3 ± 0.7**	62.6 ± 0.4	**58.9 ± 1.2**	60.3 ± 4.5	75.7 ± 1.6	**76.3 ± 1.4**	**71.7 ± 4.0**	**69.1**
Pre-trained GIN	SingleCLF	68.7 ± 1.3	78.1 ± 0.6	65.7 ± 0.6	**67.7 ± 0.8**	72.6 ± 1.5	81.3 ± 2.1	79.9 ± 0.7	84.5 ± 0.7	74.2
	MoE	67.2 ± 2.3	78.3 ± 0.5	65.4 ± 0.4	61.9 ± 0.9	73.2 ± 2.2	80.8 ± 1.0	80.7 ± 0.5	84.7 ± 0.4	74.0
	E-Ensemble	67.5 ± 2.8	78.3 ± 0.3	65.9 ± 0.2	62.3 ± 0.6	71.6 ± 1.0	80.4 ± 1.0	80.6 ± 0.5	84.3 ± 0.8	73.9
	GraphDIVE	69.1 ± 0.7	78.3 ± 0.3	66.0 ± 0.3	62.4 ± 0.7	72.1 ± 3.9	**82.6 ± 1.2**	80.1 ± 0.5	77.8 ± 2.8	73.2
	TopExpert	**69.2 ± 2.5**	**78.6 ± 0.3**	**66.2 ± 0.5**	62.9 ± 0.7	**74.1 ± 2.1**	80.3 ± 0.7	**80.8 ± 0.6**	**84.9 ± 0.8**	**74.9**

Table 1: Macro ROC-AUC on molecular property prediction tasks. The average score of 8 datasets is reported in the rightmost column. The best score for each dataset under each backbone GNN architecture is marked in bold face.

Training Details and Hyper-parameters. Each model is trained at most 200 epochs, and the training process is terminated when the validation ROC-AUC does not increase for 50 successive epochs. We train all models ten times with different random seeds and report the average score with its standard deviation. We search for the best hyper-parameter configuration through a grid search based on the validation ROC-AUC. The number of experts is chosen from $K \in \{3, 5, 7, 10\}$, and the loss balancing parameters are selected from $\alpha, \beta \in \{5, 1, 0.1, 0.01\}$. For the temperature annealing of Gumbel-Softmax, the initial temperature T_0 is set to 10, and the final temperature T_E is chosen from $\{0.01, 0.1, 1\}$. Further details are provided in Appendix A.

Quantitative Analysis (RQ1)

Table 1 reports ROC-AUC scores of each model on 8 benchmark datasets. The total average scores across 8 datasets indicate the overall capability of predicting molecular properties. We observe that TopExpert consistently shows higher performance than other baselines regardless of its backbone GNN architectures. Specifically, it achieves the highest average accuracy for all backbone GNNs, yielding higher accuracy than each of the baselines for 35 out of 40 cases on average. Further, its superior performance to the other multiple experts-based methods shows the effectiveness of our gating module that reflects enriched topological information. In the

Figure 4: Evaluation setup for robustness analysis. Example molecules of each group (Left). Splitting criteria based on the number of rings and aromatic rings (Right).

case of the Toxcast dataset with 617 tasks, which makes the target prediction much more challenging, TopExpert shows marginal improvement or even negative effects over Singe-CLF; however, its performance drop is less severe than the other multiple experts-based methods. In particular, Graph-Dive with a single gate does not perform well in datasets with multiple tasks, as discussed in Hu et al. 2021.

Robustness Analysis (RQ1 & RQ2)

Although the scaffold splitting protocol ensures that there is no common scaffold between the training and testing set, it does not consider high-level topological semantics beyond the scaffolds. For a more thorough evaluation of

Test data		SingleCLF	Expert-explicit		TopExpert	
		ROC.	ROC.	Gain	ROC.	Gain
Tox21	ID (D-1/2/3)	77.3	69.8	-7.5	77.5	+0.2
	OOD (O-1)	68.1	68.1	0.0	68.8	+0.7
	OOD (O-2)	63.9	65.5	+1.6	64.8	+0.9
HIV	ID (D-1/2/3)	67.1	60.6	-6.5	67.2	+0.1
	OOD (O-1)	58.3	63.9	+5.6	63.1	+4.8
	OOD (O-2)	63.9	61.2	-2.7	65.1	+1.2

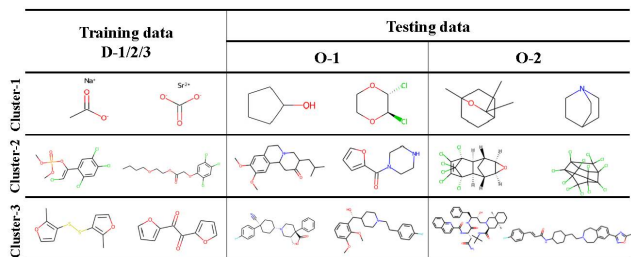Table 2: Macro ROC-AUC on ID and OOD groups (Backbone: GIN). Gain denotes the improvement over SingleCLF.

Figure 5: The molecule closest to the three cluster centroids for each molecule group, identified by TopExpert.

$\mathcal{L}_{cluster}$	$\mathcal{L}_{align}$	Tox21		HIV		BACE	
		NMI	ROC.	NMI	ROC.	NMI	ROC.
		0.096	74.5	0.149	71.4	0.322	68.8
✓		0.223	75.1	0.347	75.7	0.372	71.0
	✓	0.098	74.6	0.154	72.5	0.335	68.7
✓	✓	0.228	**75.3**	0.357	**76.3**	0.383	**71.7**

Table 3: Macro ROC-AUC of ablations (Backbone: GIN).

ful in real-world applications. In practice, the assignment ratio of training molecule groups (D-1/2/3) to each cluster (Appendix D) varies depending on the dataset, indicating that TopExpert adjusts how much it incorporates topology information by itself. Figure 5 also illustrates that TopExpert assigns molecules with unseen structures (i.e., O-1 and O-2) to the most topologically similar clusters.

Ablation Study (RQ3)

To examine how each component of our model affects the final performance, we conduct an ablation study. Note that without $\mathcal{L}_{cluster}$ and $\mathcal{L}_{align}$, our method is equivalent to the basic MoE. Table 3 shows that the clustering loss helps improve the performances over MoE. Furthermore, the best performance is achieved by using both $\mathcal{L}_{cluster}$ and $\mathcal{L}_{align}$. Besides, we report normalized mutual information (NMI) (Strehl and Ghosh 2002) to evaluate how many molecules with the same scaffold belong to the same cluster. We observe that both losses indeed improve NMI, but the extent of the growth varies widely across datasets.

Further Analysis

Additional cost for multiple experts. We compare computational costs of TopExpert with that of SingleCLF on BBBP dataset. All computations required for multiple experts are executed in parallel by GPU processors. As a result, TopExpert shows only 4% increase in training cost and 2% increase in inference cost. In terms of model size, TopExpert additionally uses 1% of the total parameters of the GIN-based model. Taken together, we conclude that the additional cost for multiple experts in TopExpert is marginal.

Parameter analysis. The effects of the loss balancing parameters (i.e., α and β) and the number of experts (i.e., K) are reported in Appendix E and F, respectively. In short, their optimal values vary depending on the dataset, and they can be easily tuned by using the validation set.

Conclusion

TopExpert adopts a clustering-based gating module that assigns input molecules into topological groups and optimizes it for effective clustering in two self-supervised ways: 1) cluster cohesion strengthening and 2) cluster-scaffold alignment. Our extensive experiments demonstrate that TopExpert significantly boosts the performance with various backbone GNN architectures, implying that topology-specific experts can offer complementary information for better generalization on test molecules with new topological patterns.

the robustness of new molecular structures having topological semantics distinguishable from the training set, we introduce a new splitting protocol. Specifically, we divide molecules into five groups, each of which has a disjoint scaffold set, according to the combination of the number of rings and aromatic rings[4] (Figure 4). We categorize D-1, D-2, and D-3 as in-distribution (ID) and O-1 and O-2 as out-of-distribution (OOD). Then, we train models using ID molecules only and evaluate the performance using both ID and OOD molecules. To further ascertain the effectiveness of our gating module, we compare TopExpert with **Expert-explicit**, a variant that explicitly and deterministically assigns an expert for each ID group (i.e., D-1, D-2, and D-3).[5]

Table 2 shows the prediction performances on ID and OOD groups. Compared to SingleCLF, TopExpert achieves comparable performance for ID groups and consistently higher performance for OOD groups, strongly supporting the TopExpert's robustness to unseen molecular structures. However, Expert-explicit shows a significant performance drop for ID groups and degraded performance for an OOD group on the HIV dataset. That is, the molecule segregation based on pre-defined rules leads to suboptimal results because it cannot capture information useful for molecular property prediction. In contrast, TopExpert clusters molecules in favor of predicting their molecular property during optimization, eventually assigning molecules to the experts in a data-driven and task-specific way, which is use-

---

[4]It is known that the number of rings is an important descriptor of molecular properties (Debnath et al. 1991).

[5]To identify the most relevant expert to each OOD molecule, Expert-explicit utilizes the centroid of each ID group. We set the number of experts to 3 for both Expert-explicit and TopExpert.

Acknowledgments

References

Ahmed, K.; Baig, M. H.; and Torresani, L. 2016. Network of experts for large-scale image categorization. In *European Conference on Computer Vision*, 516–532. Springer.

Baek, J.; Kang, M.; and Hwang, S. J. 2021. Accurate learning of graph representations with graph multiset pooling. *arXiv preprint arXiv:2102.11533*.

Balın, M. F.; Abid, A.; and Zou, J. 2019. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, 444–453. PMLR.

Bemis, G. W.; and Murcko, M. A. 1996. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry*, 39(15): 2887–2893.

Brockschmidt, M. 2020. Gnn-film: Graph neural networks with feature-wise linear modulation. In *International Conference on Machine Learning*, 1144–1152. PMLR.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J.; and Hansch, C. 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2): 786–797.

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

Fey, M.; Yuen, J.-G.; and Weichert, F. 2020. Hierarchical inter-message passing for learning on molecular graphs. *arXiv preprint arXiv:2006.12179*.

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

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; and Tkatchenko, A. 2015. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters*, 6(12): 2326–2331.

Hu, F.; Wang, L.; Wu, S.; Wang, L.; and Tan, T. 2021. Graph classification by mixture of diverse experts. *arXiv preprint arXiv:2103.15622*.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2020. Strategies For Pre-training Graph Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Hu, Y.; Stumpfe, D.; and Bajorath, J. 2016. Computational exploration of molecular scaffolds in medicinal chemistry: Miniperspective. *Journal of medicinal chemistry*, 59(9): 4062–4076.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jang, E.; Gu, S.; and Poole, B. 2017. CATEGORICAL REPARAMETERIZATION WITH GUMBEL-SOFTMAX. *stat*, 1050: 5.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kang, S.; Hwang, J.; Kweon, W.; and Yu, H. 2020. DE-RRD: A knowledge distillation framework for recommender system. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 605–614.

Kang, S.; Hwang, J.; Kweon, W.; and Yu, H. 2021. Topology distillation for recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 829–839.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Klicpera, J.; Groß, J.; and Günnemann, S. 2020. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*.

Lee, D.; Kim, S.; Lee, S.; Park, C.; and Yu, H. 2021. Learnable structural semantic readout for graph classification. In *2021 IEEE International Conference on Data Mining (ICDM)*, 1180–1185. IEEE.

Lee, S.; Lee, D.; Jang, S.; and Yu, H. 2022. Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning. *arXiv preprint arXiv:2202.13196*.

Liu, Y.; Yuan, H.; Cai, L.; and Ji, S. 2020. Deep learning of high-order interactions for protein interface prediction. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 679–687.

Lo, Y.-C.; and Torres, J. Z. 2016. Chemical similarity networks for drug discovery. *Special Topics in Drug Discovery*, 53.

Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; and He, L. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1052–1060.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the*

24th ACM SIGKDD international conference on knowledge discovery & data mining, 1930–1939.

Ma, T.; and Chen, J. 2021. Unsupervised learning of graph hierarchical abstractions with differentiable coarsening and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8856–8864.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 666–704.

Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; and Bajorath, J. 2019. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS omega*, 4(1): 1027–1032.

Peng, H.; Li, J.; Gong, Q.; Ning, Y.; Wang, S.; and He, L. 2020. Motif-matching based subgraph-level attentional convolutional network for graph classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5387–5394.

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754.

Rouvray, D. H. 1986. Predicting chemistry from topology. *Scientific American*, 255(3): 40–47.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shen, T.; Ott, M.; Auli, M.; and Ranzato, M. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, 5719–5728. PMLR.

Strehl, A.; and Ghosh, J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec): 583–617.

Subramonian, A. 2021. Motif-driven contrastive learning of graph representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15980–15981.

Sun, F.-Y.; Hoffmann, J.; Verma, V.; and Tang, J. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. In *ICLR*.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How powerful are graph neural networks? In *ICLR*.

Yang, H.; Ma, K.; and Cheng, J. 2021. Rethinking graph regularization for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4573–4581.

Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388.

Yang, Y.; Feng, Z.; Song, M.; and Wang, X. 2020. Factorizable graph convolutional networks. *Advances in Neural Information Processing Systems*, 33: 20286–20296.

Yu, Z.; and Gao, H. 2022. Molecular Representation Learning via Heterogeneous Motif Graph Neural Networks. In *International Conference on Machine Learning*, 25581–25594. PMLR.