Pre-training Graph Neural Networks for Molecular Representations: Retrospect and Prospect

Anonymous Authors[1]

Abstract

Recent years have witnessed remarkable advances in molecular representation learning using Graph Neural Networks (GNNs). To fully exploit the unlabeled molecular data, researchers first pre-train GNNs on large-scale molecular databases and then fine-tune these pre-trained Graph Models (GMs) in downstream tasks. The knowledge implicitly encoded in model parameters can benefit various downstream tasks and help to alleviate several fundamental challenges of molecular representation learning. In this paper, we provide a comprehensive survey of pre-trained GMs for molecular representations. We first briefly present the limitations of molecular graph representation learning and thus introduce the motivation for molecular graph pre-training. Next, we systematically categorize existing pre-trained GMs based on a taxonomy from four different perspectives including model architectures, pre-training strategies, tuning strategies, and applications. Finally, we outline several promising research directions that can serve as a guideline for future studies.

1. Backgrounds

GNNs have gained popularity in various molecule-related tasks for their ability of modeling structural information such as molecular graphs. However, two fundamental challenges impede the wider usage of existing supervised graph learning on molecular graphs: (1)*Scarce Labeled Data:* Task-specific labeled molecules can be extremely scarce because high-quality data labeling for molecular graphs often requires time-consuming and resource-costly wet-lab experiments. (2) *Out-of-distribution Generalization:* Existing GNNs lack out-of-distribution generalization abilities so that their performance substantially degrades when there exist distribution shifts between training and testing molecular graphs. This issue is common to see in real-world applications such as predicting the properties of a brand-new, just synthesized molecule, which is different from all molecules that have been synthesized so far (Hu* et al., 2020).

Indeed, nearly all the deep learning domains are confronted with these challenges. To alleviate these issues, certain progress has been made. For example, the *pretrain-then-finetune* paradigm of pre-trained Language Models (LMs) is thriving in Natural Language Processing (NLP) community. Specifically, they first pre-train the models on a large-scale corpus and then fine-tune these models in various downstream tasks. With the emergence of Transformer (Vaswani et al., 2017), pre-trained LMs such as BERT (Devlin et al., 2019) have become dominative roles for NLP, which have established state-of-the-art results for various NLP tasks.

Inspired by the proliferation, tremendous efforts have been devoted to pre-trained Graph Models (GMs) recently. It is widely recognized that the well pre-trained GMs can provide a better initial point across downstream tasks and leads to wider optima with better generalization than training from scratch (Hao et al., 2019) in a scarce data regime. In this paper, we provide researchers with synthesis and pointer to related research on molecular GMs. Existing surveys related to this area have only partially focused on self-supervised learning on graphs (Liu et al., 2021; Xie et al., 2021), but did not go broader to the other important ingredients such as supervised pre-training, tuning strategies, various extensions, and their applications in molecular representations. Overall, the contributions can be summarized as follows: **(1)** *Comprehensive review.* Our survey serves as a pioneering work that presents a comprehensive review of pre-trained GMs for molecular representations. **(2)***New taxonomy.* We propose a new taxonomy shown in Figure 1, which categorizes existing works from the following perspectives: Model architectures; Pre-training strategies; Tuning Strategies; Applications in molecular representations. **(3)** *Abundant resources.* We collect abundant resources on this topic, including open-sourced pre-trained GMs, pre-training datasets, paper lists and etc. We will release these resources upon acceptance. **(4)** *Future directions.* We discuss the limitations of existing works and suggest possible future research directions.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

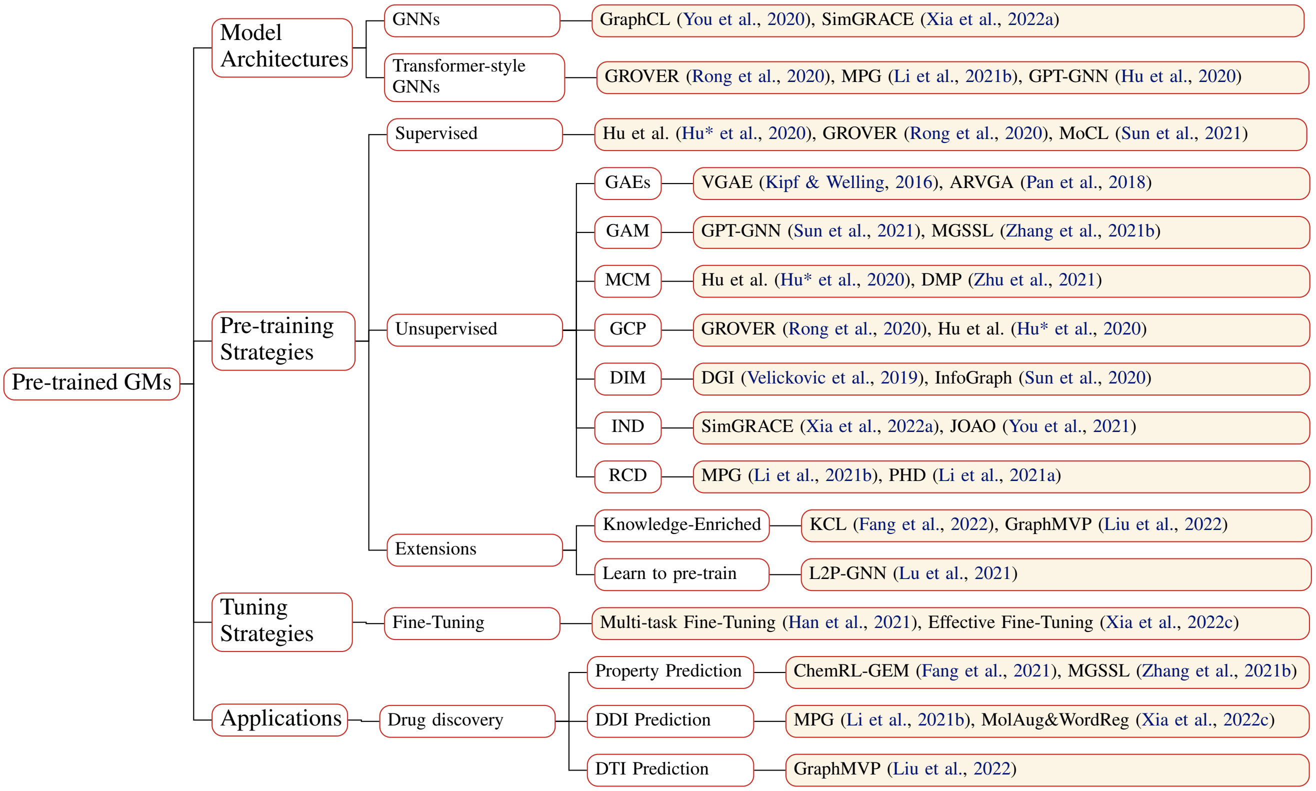
Pre-trained GMs
Model Architectures
GNNs
GraphCL (You et al., 2020), SimGRACE (Xia et al., 2022a)
Transformer-style GNNs
GROVER (Rong et al., 2020), MPG (Li et al., 2021b), GPT-GNN (Hu et al., 2020)
Pre-training Strategies
Supervised
Hu et al. (Hu* et al., 2020), GROVER (Rong et al., 2020), MoCL (Sun et al., 2021)
Unsupervised
GAEs
VGAE (Kipf & Welling, 2016), ARVGA (Pan et al., 2018)
GAM
GPT-GNN (Sun et al., 2021), MGSSL (Zhang et al., 2021b)
MCM
Hu et al. (Hu* et al., 2020), DMP (Zhu et al., 2021)
GCP
GROVER (Rong et al., 2020), Hu et al. (Hu* et al., 2020)
DIM
DGI (Velickovic et al., 2019), InfoGraph (Sun et al., 2020)
IND
SimGRACE (Xia et al., 2022a), JOAO (You et al., 2021)
RCD
MPG (Li et al., 2021b), PHD (Li et al., 2021a)
Extensions
Knowledge-Enriched
KCL (Fang et al., 2022), GraphMVP (Liu et al., 2022)
Learn to pre-train
L2P-GNN (Lu et al., 2021)
Tuning Strategies
Fine-Tuning
Multi-task Fine-Tuning (Han et al., 2021), Effective Fine-Tuning (Xia et al., 2022c)
Applications
Drug discovery
Property Prediction
ChemRL-GEM (Fang et al., 2021), MGSSL (Zhang et al., 2021b)
DDI Prediction
MPG (Li et al., 2021b), MolAug&WordReg (Xia et al., 2022c)
DTI Prediction
GraphMVP (Liu et al., 2022)

Figure 1. Taxonomy of pre-trained GMs with representative examples.

2. Overview of pre-trained GMs

With the development of computational power and the constant enhancement of training skills, as demonstrated in Figure 2, recent pre-trained GMs have embraced a transfer learning (Pan & Yang, 2009) setting where the goal is to pre-train a generic encoder that can deal with various molecule-related tasks. Initially, Hu et al. (Hu* et al., 2020) initialize a 5-layer Graph Isomorphism Network (GIN) (Xu et al., 2019) with the pre-trained GM obtained with diverse pre-training tasks to capture both atom-level and molecule-level information. Inspired by this precursor, the modern pre-trained GMs are usually trained with larger-scale databases, more powerful or deeper architectures (e.g., the hybrid of GNNs and Transformer (Vaswani et al., 2017) we describe below), and more advanced pre-training strategies. For example, the pre-trained GMs like GROVER (Rong et al., 2020) and MPG (Li et al., 2021b) with huge parameters have shown their powerful ability in learning universal molecular graph representations.

3. Model Architectures

The model architectures of pre-trained GMs broadly fall into two categories: Graph Neural Networks (GNNs), Transformer-style GNNs. We elaborate on them below.

3.1. Graph Neural Networks (GNNs)

GNNs have emerged as the dominant tools for modeling graph data. The structure of graph data guides the aggregation of local neighborhood information and leads to a more contextual representation for each node. Also, we can adopt a graph pooling operation (Mesquita et al., 2020) to get the representation for the whole graph. Molecules can naturally be modeled as graph data with their atoms as nodes and chemical bonds as edges. For pre-trained GMs in molecular representation learning, GIN (Xu et al., 2019) is the most popular encoder for its powerful expressiveness. The other common GNNs such as Graph Convolutional Network (GCN) (Kipf & Welling, 2017) and GraphSAGE (Hamilton et al., 2017) can also serve as the encoder for pre-training, whereas their performance is often inferior to GIN. Additionally, as revealed in a recent study (Hu* et al., 2020), pre-training Graph Attention Network (GAT) (Velickovic et al., 2018) will incur dramatical 'negative transfer', which means that the pre-training-then-finetuning paradigm with GAT falls behind training from scratch by large margins. It is promising to explore why this phenomenon would occur.

3.2. Transformer-style GNNs

Although GNNs have achieved spectacular performance in pre-training on molecular graphs, their limited parame-

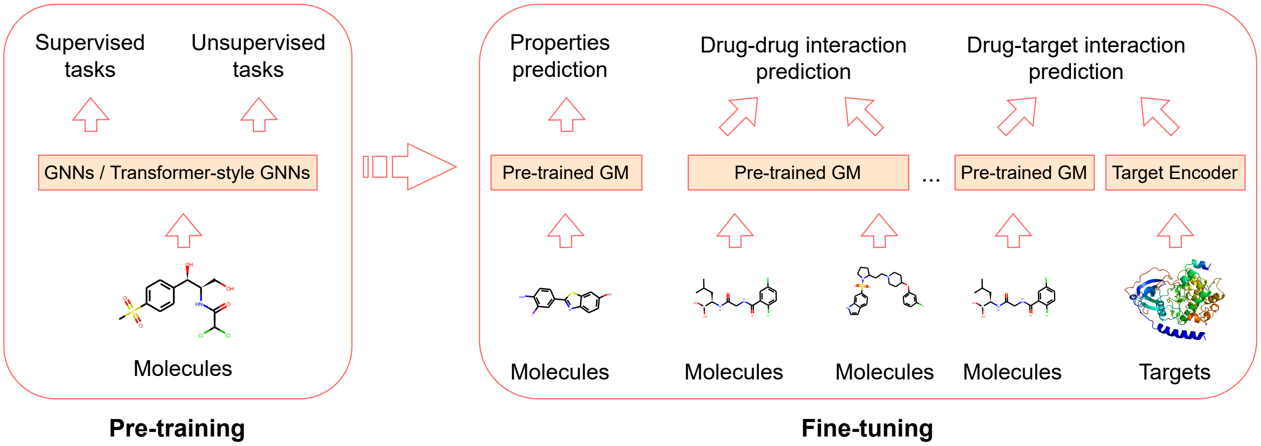

Figure 2. The pre-training and fine-tuning flow of GMs: First, a GNN is pre-trained with supervised or unsupervised tasks. Secondly, the pre-trained model and its parameters are then used to initialize models for various downstream tasks on the input molecular graphs.

ters are not enough for large-scale datasets with millions of molecules. On the other hand, Transformer has become the de facto standard for doing large-scale pre-training in NLP. Therefore, several recent works try to integrate GNNs into Transformer-style models, which we name *Transformer-style GNNs*. For example, GROVER (Rong et al., 2020) first utilizes GNNs to capture local structural information of the graph data, and then the outputs of the GNNs are regarded as queries, keys, and values for the Transformer encoder. They claim that this bi-level information extraction strategy largely enhances the representational power. Analogously, MPG (Li et al., 2021b) devises a neighbor attention module to produce a message representation for each node and feed it to a fully connected feed-forward network. With the proper message representation obtained, they adopt a Gated Recurrent Unit (GRU) network (Cho et al., 2014) to update node representation. Additionally, there are some recent works that try to incorporate the graph information into the vanilla Transformer with improved positional embedding (Hussain et al., 2021; Cai & Lam, 2020; Mialon et al., 2021) or improved attention matrix from graphs (Ying et al., 2021). Among them, Graphormer (Ying et al., 2021) is the most popular one for molecular graphs pre-training.

4. Pre-training Strategies

In this section, we will elaborate on various strategies for pre-training on molecular graphs. We summarize and formulate the above pre-training strategies using a unified symbolic system in Table 1.

4.1. Supervised Strategies

Although the supervised labels for molecules are often time-consuming and expensive to collect, some cheaper annotations that may be less related to downstream tasks can also help pre-training on molecular graphs. For example, Hu et al. (Hu* et al., 2020) propose to pre-train GNNs to predict essentially all the properties of molecules that have been experimentally measured so far. Also, they leave a future work to take the structural similarities between two molecular graphs as supervision. Inspired by this, MoCL (Sun et al., 2021) first calculates the Tanimoto coefficient (Bajusz et al., 2015) between two molecules as the measure of structural similarity, which serves as the supervision for the pre-training. Additionally, given that motifs in molecular graphs usually correspond to functional groups that are indicative of molecular properties, some recent works such as GROVER (Rong et al., 2020) and MGSSL (Zhang et al., 2021b) detect the motifs using the professional software such as RDkit[1] or developed algorithms (Ertl, 2017) and then predict the presence of the motifs or generate the motifs respectively. Although the supervised pre-training brings remarkable improvements, some supervised pre-training tasks might be unrelated to the downstream tasks of interest and can even hurt the downstream performance.

4.2. Unsupervised Strategies

4.2.1. Graph AutoEncoders (GAEs)

Graph reconstruction serves as a natural self-supervision for learning discriminative molecular graph representations. The prediction targets in graph reconstruction are certain parts of the given molecular graphs such as the attribute of a subset of atoms or chemical bonds. Inspired by the success of AutoEncoders in Computer Vision (CV), various GAEs have been developed recently. Among many, GAE (Kipf & Welling, 2016) is the simplest version of the graph autoen-

[1] https://www.rdkit.org/

Table 1. Loss functions of supervised and unsupervised pre-training strategies. $\mathcal{G}$ = $(\mathcal{V}, \mathcal{E})$: The molecular graph; $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$: Atoms set; $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$: Bonds set; $\boldsymbol{X} \in \mathbb{R}^{N \times F}$: The atom attributes matrix. F denotes the feature dimensions.

Task	**Loss Function**	**Description**
Supervised	$\mathcal{L}_{\texttt{Supervised}} = -\log p\left(\boldsymbol{Y} \mid \mathcal{G}\right)$	Supervised pre-training. $\boldsymbol{Y}$ is the given label.
GAEs	$\mathcal{L}_{\texttt{GAEs}} = -\log p\left(\boldsymbol{X}, \mathcal{E} \mid \mathcal{G}\right)$	Graph reconstruction.
GAM	$\mathcal{L}_{\texttt{GAM}} = -\sum_{i=1}^{\lvert\mathcal{V}\rvert} \log p\left(\boldsymbol{X}_i, \mathcal{E}_i \mid \boldsymbol{X}_{<i}, \mathcal{E}_{<i}\right)$	$\boldsymbol{X}_{<i}, \mathcal{E}_{<i}$ are the attributes and edges generated before node i respectively.
MCM	$\mathcal{L}_{\texttt{MCM}} = -\sum_{\widehat{\mathcal{G}} \in m(\mathcal{G})} \log p\left(\widehat{\mathcal{G}} \mid \mathcal{G}_{\backslash m(\mathcal{G})}\right)$	$m(\mathcal{G})$ are the masked components from $\mathcal{G}$ and $\mathcal{G}_{\backslash m(\mathcal{G})}$ are the rest.
GCP	$\mathcal{L}_{\texttt{GCP}} = -\log p(t \mid \mathcal{G}_1, \mathcal{G}_2)$	$t = 1$ if neighborhood graph $\mathcal{G}_1$ and contexts $\mathcal{G}_2$ belong to the same node.
IND	$\mathcal{L}_{\texttt{IND}} = -s\left(\mathcal{G}, \mathcal{G}^+\right) + \log \sum_{\mathcal{G}^- \in \mathcal{N}} s\left(\mathcal{G}, \mathcal{G}^-\right)$	$\mathcal{N}$ is a set of negatives; $\mathcal{G}^+$ is a positive sample.
DIM	$\mathcal{L}_{\texttt{IND}} = -s\left(\mathcal{G}, \mathcal{C}\right) + \log \sum_{\mathcal{C}^- \in \mathcal{N}} s\left(\mathcal{G}, \mathcal{C}^-\right)$	$\mathcal{N}$ is a set of negatives; $\mathcal{C}$ is a substructure of $\mathcal{G}$.
RCD	$\mathcal{L}_{\texttt{RCD}} = -\log p(t \mid \mathcal{G}_1, \mathcal{G}_2)$	$t = 1$ if two half graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are homologous couples.

coders, which reconstructs the adjacency matrix of the original graph using the binary cross-entropy loss. Also, there exist multiple variants of GAEs that utilize graph reconstruction to pre-train the GNNs. Representative examples include VGAE (Kipf & Welling, 2016), MGAE (Wang et al., 2017), ARVGA (Pan et al., 2018), SIG-VAE (Hasanzadeh et al., 2019) and etc. Although GAEs can learn meaningful representations for molecular graphs, they fail to capture the inter-molecule relationships, which accounts for their poorer performance.

4.2.2. Graph Autoregressive Modeling (GAM)

Following the idea of GPT (Brown et al., 2020) that conducts generative language model pre-training, GPT-GNN (Hu et al., 2020) proposes an autoregressive framework to perform reconstruction on given graphs iteratively, which is different from graph autoencoders that reconstruct the graph all at once. In particular, given a graph with its nodes and edges randomly masked, GPT-GNN generates one masked node and its edges at a time and optimizes the parameterized models via maximizing the likelihood of the node and edges generated in the current iteration. Then, it iteratively generates nodes and edges until all masked nodes are generated. Analogously, MGSSL (Zhang et al., 2021b) generates molecular graph motifs in an autoregressive way based on existing motifs and connections. Compared with other pre-training strategies, the pre-trained GMs obtained with GAM are better at molecular graph generation. However, autoregressive generation incurs heavy computationally overhead, which impedes wider usage in large-scale pre-training on molecular graphs.

4.2.3. Masked Components Modeling (MCM)

Similar to masked language modeling (MLM) which masks out some tokens from the input sentences and then trains the model to predict the masked tokens by the rest of the tokens (Devlin et al., 2019), MCM first masks out some components (e.g., atoms, bonds, subgraphs and etc.) of the molecular graphs and then trains the model to predict them. For example, Hu et.al (Hu* et al., 2020) propose attribute masking where the input atom/chemical bond attributes are randomly masked, and the GNN is asked to predict them. Also, GROVER (Rong et al., 2020) tries to predict the masked subgraphs to capture the contextual information in molecular graphs. These masking methods are especially beneficial for richly-annotated molecular graphs. For example, masking node attributes (atom type) enables GNNs to learn simple chemistry rules such as valency, as well as potentially more complex chemistry phenomena such as the electronic or steric properties of functional groups. Additionally, compared with GAM we describe in section 4.2.2, MCM predicts the masked components (atoms/bonds) based on their surrounding environments while GAM predicts them only dependent on the components appearing before them in the handcrafted sequence. As a result, MCM allows the pre-trained GMs to better capture chemical rules. However, the input to pre-training GNNs in MCM contains artificial symbols that never occur in downstream tasks, which creates a pretrain-finetune discrepancy (Hu* et al., 2020; Yang et al., 2019). This issue remains unsolved in molecular graph pre-training.

4.2.4. Graph Context Prediction (GCP)

GCP is proposed to explore the distribution of graph structure in molecular graph data. For example, Hu et al. (Hu* et al., 2020) use subgraphs in molecules to predict their surrounding graph structures. They pre-train a GNN so that it maps atoms appearing in similar structural contexts to nearby embeddings. GROVER tries to predict the context-aware properties of the target atom/bond within some local subgraph. Here, the properties refer to some atom-bond count terms around the target atom/bond. Although effective, GCP requires an auxiliary GNN to encode the context into a fixed vector, which is redundant for large-scale pre-training.

4.2.5. Graph Contrastive Learning (GCL)

Graph contrastive pre-training has emerged as the most popular strategy for molecular representations, which broadly fall into two categories based on their contrastive granularities (e.g., atom vs. molecule or molecule vs. molecule): *Deep InfoMax* and *Instance Discrimination*.

Deep InfoMax (DIM) Deep InfoMax is originally proposed for images, which improves the quality of the representation by maximizing the mutual information between an image representation and local regions of the image (Hjelm et al., 2019). For molecular graphs, initially, InfoGraph (Sun et al., 2020) is proposed to obtain expressive representations for molecules or atoms via maximizing the mutual information between graph-level representations and substructure-level representations of different granularity. Similarly, MVGRL (Hassani & Khasahmadi, 2020) performs node diffusion to generate an augmented view and then maximizes the mutual information between original and augmented views by contrasting atom representations of one view with molecule representations of the other view and vice versa.

Instance Discrimination (IND) IND is one of the most popular pre-training strategies which embeds augmented versions of the anchor molecular graph close to each other (positive pairs) and pushes the embeddings of other molecular graphs (negative pairs) apart. For molecular representations, GraphCL (You et al., 2020) and its variants (You et al., 2021; Sun et al., 2021; Suresh et al., 2021; Fang et al., 2022; Xu et al., 2021a) propose various advanced augmentations strategies for graph-level pre-training. More recently, some works such as BGRL (T. et al., 2021), CCA-SSG (Zhang et al., 2021a), LP-Info (You et al., 2022) and SimGRACE (Xia et al., 2022a) try to simplify graph contrastive pre-training via discarding the negatives, parameterized mutual information estimator, or even molecular graph data augmentations, respectively.

Although molecular graph contrastive pre-training has achieved spectacular results, there are several critical issues impeding its broader applications. For example, it is difficult to preserve semantics during molecular graph augmentations. Existing solutions picking augmentations with manual trial-and-errors (You et al., 2020), cumbersome optimization (You et al., 2021) or the guidance of expensive domain knowledge (Sun et al., 2021; Xia et al., 2021) are unsatisfactory. It remains to be explored whether there are more suitable augmentations for molecular graphs. On the other hand, we push away all the other molecular graphs regardless of their true semantics in graph contrastive pre-training, which will undesirably push away the molecules of similar properties as advocated in a recent work (Xia et al., 2022b).

4.2.6. Replaced Component Detection (RCD)

To capture the global information of molecular graphs, RCD is proposed as a graph-level pre-training task on a random permutation of input molecular graphs. For example, PHD (Li et al., 2021a) first decomposes each molecular graph in the database into two half-graphs and replaces one of them with a half-graph from other molecular graphs randomly. The GNN encoder is pre-trained to detect whether two half-graphs are homologous couples. Although RCD can help GMs capture intrinsic patterns underlying the graph structures, it remains a binary classification task in essence, which is less challenging than MCM we elaborate on in section 4.2.3.

4.3. Extensions

4.3.1. Knowledge-Enriched Pre-training

Pre-trained GMs usually learn universal molecular graph representation from the general-purpose molecular database. However, they often lack domain-specific knowledge. To enhance their performance, several recent works try to inject external knowledge during pre-training. For example, GraphCL (You et al., 2020) first points out that bond perturbation is conceptually incompatible with domain knowledge and empirically unhelpful for downstream performance for chemical compounds. Therefore, they avoid adopting bond perturbation for molecular graph augmentation. To incorporate the domain knowledge into pre-training more explicitly, MoCL (Sun et al., 2021) proposed a new molecular augmentation operator called substructure substitution, in which a valid substructure of a molecule is replaced by a bioisostere (Meanwell, 2011) which produces a new molecule with similar physical or chemical properties as the original one. They compile 230 substitution rules from domain resources in total and empirically validate their effectiveness. More recently, to capture the correlations between atoms that have common attributes but are not directly connected by bonds, KCL (Fang et al., 2022) construct a chemical element Knowledge Graph (KG) to summarize microscopic associations between elements and propose a novel Knowledge-enhanced Contrastive Learning (KCL) framework for molecular representation learning. Considering that 3D geometric information of molecules also plays a vital role in predicting molecular functionalities, 3DInfoMax (Stärk et al., 2021) proposes pre-training a model to reason about the geometry of molecules given only their 2D molecular graphs while GraphMVP (Liu et al., 2022) performs self-supervised pre-training via maximizing the correspondence and consistency between 2D topological structures and 3D geometric views. Additionally, ChemRL-GEM (Fang et al., 2021) proposes to utilize the molecular geometry information to enhance molecular representation learning. They design a geometry-based graph neural net-

Table 2. List of Representative and open-sourced pre-trained GMs. KG: Chemical Element Knowledge Graph.

pre-trained GMs	Input	Architecture	Pre-Training Task	Pre-training Database	# Params.
Hu et al. (Hu* et al., 2020)	Graph	5-layer GIN	GCP + MCM	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
GraphCL (You et al., 2020)	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
JOAO (You et al., 2021)	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
AD-GCL (Suresh et al., 2021)	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
GraphLog (Xu et al., 2021b)	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
GROVER (Rong et al., 2020)	Graph	GTransformer (Rong et al., 2020)	GCP + MCM	ZINC + ChEMBL (10M)	48M$\sim$100M
MGSSL (Zhang et al., 2021b)	Graph	5-layer GIN	MCM + GAM	ZINC15 (250K)	$\sim$ 2M
PMG (Li et al., 2021b)	Graph	MolGNet (Li et al., 2021b)	RCD + MCM	ZINC + ChEMBL (11M)	53M
LP-Info (You et al., 2022)	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
SimGRACE (Xia et al., 2022a)	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	$\sim$ 2M
MolCLR (Wang et al., 2021)	Graph + SMILES	GCN + GIN	IND	PubChem (10M)	N/A
DMP (Zhu et al., 2021)	Graph + SMILES	DeeperGCN + Transformer	MCM + IND	PubChem (110M)	104.1 M
ChemRL-GEM (Fang et al., 2021)	Graph + Geometry	GeoGNN (Fang et al., 2021)	MCM+GCP	ZINC15 (20M)	N/A
KCL (Fang et al., 2022)	Graph + KG	GCN + KMPNN (Fang et al., 2022)	IND	ZINC15 (250K)	<1M
3D Infomax (Stärk et al., 2021)	2D and 3D molecule	PNA (Corso et al., 2020)	IND	QM9(50K) + GEOM(140K) + QMugs(620K)	N/A
GraphMVP (Liu et al., 2022)	2D and 3D molecule	5-layer GIN + SchNet (Schütt et al., 2017)	IND + GAEs	GEOM (50k)	$\sim$ 2M

work architecture as well as several geometry-level self-supervised learning strategies (the bond lengths prediction, the bond angles prediction, and the atomic distance matrices prediction) to capture the molecular geometry knowledge during pre-training. Analogously, GeomGCL (Li et al., 2021c) regards 2D and 3D views of the same molecule as positive pairs while the remaining pairs as negative pairs for contrastive learning. In this way, they can avoid the random augmentation process of molecular graphs in contrastive molecular graph pre-training. Although knowledge-enhanced pre-training help GMs capture chemical domain knowledge, however, it requires expensive domain knowledge as guidance, which poses a hurdle to broader applications.

4.3.2. Learn to Pre-train

Due to the divergence of the optimization objectives between pre-training and fine-tuning steps, there exists a gap between them which will significantly hurt the generalization ability of pre-trained molecular graph models. To narrow this gap, L2P-GNN (Lu et al., 2021) simulates the fine-tuning via creating new tasks during pre-training. This setup enables pre-trained GMs to adapt to new tasks quickly and leads to better generalization on downstream tasks.

5. Tuning Strategies

Although multiple pre-trained GMs are open-sourced for public usage (as listed in Table 2), the process of vanilla fine-tuning them is still brittle. For example, Xia et al. (Xia et al., 2022c) observe that pre-trained GMs are prone to over-fit insufficient labeled molecules for downstream tasks due to their high complexity. In particular, unlike image or text data, getting labels for biochemical graph data often requires laborious wet-lab experiments. To enrich the labeled data of downstream tasks, they propose to augment molecular graph data with chemical enantiomers and homologies, which share the similar physical (permeability, solubility and etc.) or chemical (toxicity, side effect, and etc.) properties with original molecules. To control the complexity of pre-trained GMs, they introduce a new regularization built on dropout which encourages the output of GMs not to change much when injecting a small perturbation and thus effectively controls GMs' capacity. Additionally, catastrophic forgetting often happens when adapting pre-trained GMs to downstream tasks. Namely, pre-trained GMs often forget their learned general knowledge when fine-tuning. To alleviate this issue, Han et al. (Han et al., 2021) utilize meta learning (Hospedales et al., 2021) to adaptively select and combine various pre-training tasks with the target task in fine-tuning stage to achieve a better adaptation. This preserves sufficient knowledge captured by self-supervised pre-training tasks while improving the effectiveness of transfer learning on GNNs. However, it assumes the pre-training tasks of pre-training are available, which is impractical because the pre-training tasks are often unknown to users in the downstream tasks. As an alternative, GTOT-Tuning (Zhang et al., 2022) introduces *GTOT Regularizer*, which can utilize graph structure to preserve the local feature invariances between finetuned and pre-trained models and thus alleviate the catastrophic forgetting issue.

6. Applications

Recently, the advancements in pre-training on molecular graphs provide opportunities to expedite drug discovery and development pipeline. In this section, we demonstrate several promising application scenarios that can embrace the power of pre-trained graph models.

6.1. Molecular Property Prediction (MPP)

In practice, the oral bioavailability of a brand-new drug is related to many properties, such as solubility in the gastrointestinal tract, intestinal membrane permeability, and

Table 3. Summary for the most widely-used chemical datasets for evaluating pre-trained GMs.

Dataset	Task	# Tasks	# Molecules	# Proteins	# Molecule-Protein	# Molecule-Molecule
BBBP	MPP (Classification)	1	2,039	–	–	–
Tox21	MPP (Classification)	12	7,831	–	–	–
ToxCast	MPP (Classification)	617	8,576	–	–	–
Sider	MPP (Classification)	27	1,427	–	–	–
ClinTox	MPP (Classification)	2	1,478	–	–	–
MUV	MPP (Classification)	17	93,087	–	–	–
HIV	MPP (Classification)	1	41,127	–	–	–
Bace	MPP (Classification)	1	1,513	–	–	–
ESOL	MPP (Regression)	1	1,128	–	–	–
FreeSolv	MPP (Regression)	1	643	–	–	–
Lipophilicity	MPP (Regression)	1	4,200	–	–	–
TWOSIDES	DDI (Classification)	1	3,300	–	–	63,000
DeepDDI	DDI (Classification)	1	192,284	–	–	19,187
Davis	DTI (Regression)	1	68	379	30,056	–
KIBA	DTI (Regression)	1	2,068	229	118,254	–
C. Elegans	DTI (Regression)	1	1,434	2,504	4,000 (positive interactions)	–
Human	DTI (Regression)	1	1,502	852	3,369 (positive interactions)	–

intestinal/hepatic first-pass metabolism (Hou et al., 2007). However, it is often laborious and even unsafe to conduct such experiments on human bodies. As an alternative, pre-trained GMs can capture abundant knowledge from the unlabeled molecules and can be directly applied as a molecule encoder to obtain expressive representations for the new drug (Wang et al., 2021; Rong et al., 2020), which is conducive to molecular property prediction.

MoleculeNet (Wu et al., 2018) is the most common benchmark for molecular property prediction, which includes 700, 000 molecules from PubChem (Kim et al., 2016), PubChem BioAssasy (Wang et al., 2012) and ChEMBL (Gaulton et al., 2012). The properties of molecules broadly fall into four categories: physiological, biophysical, physicochemical, and quantum mechanics. Additionally, there are 17 datasets in MoleculeNet in total, among which FreeSolv, ESOL, MUV, HIV, BACE, BBBP, Tox21, ToxCast, SIDER and Clintox are the most commonly used ones to evaluate pre-trained GMs. The molecular property prediction using MoleculeNet can be regarded as multi-label binary classification or regression tasks in machine learning. We summarize the most widely-used datasets for evaluating pre-trained GMs in Table 3.

6.2. Drug-Drug Interaction (DDI)

Drug-drug interaction (DDI) prediction is also imperative in drug discovery pipelines because DDIs may lead to adverse drug reactions (ADRs) which will damage the health or even cause death. Therefore, DDI potential is an important part of drug development and regulatory investigation prior to market approval. Considering that molecules with similar structures may produce the same side effects, computer-assisted techniques predict DDT via comparing molecular structural similarity (Scheiber et al., 2009a;b). DDI prediction tasks can be regarded as a task that classifies the influence of combining drugs into three categories: synergistic, additive, and antagonistic. Works on molecular graph pre-training, such as MPG (Li et al., 2021b), and WordReg & MolAug (Xia et al., 2022c), have adopted DDI prediction as a downstream task to validate the effectiveness of the pre-trained GMs.

DDI datasets generally come from clinical observational cases. However, these datasets are often of low quality because the variables in real-world cases are often unmeasured or sparse. To remedy these drawbacks, Tatonetti et al. (Tatonetti et al., 2012) constructed a high-quality dataset TWOSIDES, which includes over 3,300 drugs and 63,000 combinations connected to millions of potential adverse reactions. Additionally, a DDI dataset was proposed in DeepDDI (Ryu et al., 2018) and extracted from DrugBank (Wishart et al., 2008), which is a comprehensive drug database crucial for computer-assisted drug discovery. The DDI dataset in DeepDDI is a multi-classification task containing 192,284 DDIs contributed by 191,878 drug pairs.

6.3. Drug-Target Interaction (DTI)

Drug-target interaction (DTI) prediction is a crucial task in drug discovery. When a new indication occurs, the best choice for coping is to recycle approved drugs because of their availability and known safety profiles. DTI prediction can reduce the need for further drug development and lower the drug safety risk. The framework of DTI (Nguyen et al., 2021) is consisted of two encoders. One is the pre-trained GM for molecules and the other is the encoder for the target (e.g., convolutional neural networks for amino acid sequence). In DTI, we aim to predict the affinity scores between the molecular drugs and protein targets. In this case, pre-trained GMs can be directly applied as a drug molecule encoder and the well pre-trained model weights can be regarded as the initial weights of the drug encoder. The drug encoder and target encoder are then trained with the DTI prediction task. Related works including GraphMVP (Liu et al., 2022), MPG (Li et al., 2021b) and WordReg & MolAug (Xia et al., 2022c) have followed this setting to achieve DTI prediction.

Human and *Caenorhabditis elegans* are two DTI datasets specific for DTI prediction task (Liu et al., 2015). The positive data of these datasets were selected from two experiment-based databases: DrugBank (Wishart et al., 2008) and Matador (Günther et al., 2007). Instead of randomly choosing compounds and proteins to construct negative samples, the negative samples of human and C. elegans were obtained through a systematic screening framework to ensure their high credibility. In the human dataset, 1,052 unique compounds and 852 unique proteins constituted 3,369 positive interactions. In C. elegans dataset, 4,000 positive interactions were found between 1,434 unique compounds and 2,504 unique proteins. Additionally, Davis (Davis et al., 2011) measures the binding affinities between kinase inhibitors and kinases with the K_d value (kinase dissociation constant). KIBA (Tang et al., 2014) contains binding affinities for kinase inhibitors from different sources, including K_i, K_d and IC_{50}. Both datasets can also be utilized to evaluate the pre-trained GMs.

7. Conclusion and Future Outlooks

Despite the fruitful progress of pre-trained GMs, challenges still exist due to the complexity of graph data. In this section, we suggest several promising research directions for the future.

7.1. Better Knowledge Transfer

Currently, tremendous efforts are focusing on pre-training strategies. However, how to leverage these pre-trained GMs is still under-explored compared to pre-trained language models in NLP. Fine-tuning is a dominant technique to adapt the knowledge to various downstream tasks, but there are several nonnegligible deficiencies to be solved. The first one is poor generalization of pre-trained GMs especially for various molecular tasks where collecting labeled data is laborious. The second issue is parameter inefficiency. The fine-tuned parameters vary across both datasets and tasks, which are often huge in scale and thus inconvenient in special scenarios such as low-capacity devices. Furthermore, there are some promising alternatives to mine the knowledge from pre-trained GMs. For example, distilling the knowledge from pre-trained GMs as adopted in NLP is expected (Yang et al., 2020).

7.2. Better Model Architectures, Tasks for Pre-training on Molecular Graphs

As revealed in Section 3, the application of powerful graph neural network architecture GAT in molecular graph pre-training will incur negative transfer issue. It is promising to explore why this phenomenon would occur and what kind of GNN architectures are most suitable for molecular graph pre-training. Additionally, for large-scale pre-training, how to integrate expressive GNNs and Transformer into a unified encoder deserves more attention. On other hand, some representative pre-training strategies are fraught with issues as we pointed out in section 4. How to mitigate these critical issues is a fruitful direction for the future.

7.3. More Reliable Benchmarks for Fair Evaluation

MoleculeNet has become the most popular benchmark for evaluating pre-trained GMs. However, this benchmark is potentially brittle because most datasets are insufficiently labeled and over-parameterized pre-trained models are prone to overfit them. Worse still, the performance of GMs on these datasets is unsteady with diverse random seeds in such a small-data regime. More Reliable benchmarks for the fair evaluation of pre-trained GMs are expected.

7.4. Interpretability of Pre-trained GMs

Despite their proliferation, a major limitation of pre-trained GMs is that they are not amenable to interpretability. Worse still, unlike CNNs for images, interpreting pre-trained GMs is more difficult due to the complexities of both the Transformer-style architecture and graph data. However, for some specific scenarios like molecular toxicity prediction, it is of vital importance for the pre-trained GMs to possess the ability to explain the reason why a molecule is non-toxic. Also, interpretability can accelerate some scientific findings such as identifying biomarkers. Overall, as a key component in graph-related applications, the interpretability of pre-trained GMs remains to be explored further in many respects, which helps us understand how pre-trained GMs work and provides a guide for better usage.

7.5. Broader Scope of Applications

Pre-trained GMs have been applied in various tasks of drug discovery. However, it remains underexplored how pre-trained GMs can benefit more tasks such as chemical reaction prediction (Schwaller et al., 2021), retrosynthesis (Segler et al., 2018), and molecule generation (Du et al., 2022). Additionally, recent works have demonstrated that GNNs can help learn expressive representations for proteins (Xia & Ku, 2021). More endeavors are expected to study whether pre-trained GMs are conducive to protein representation learning.

References

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7, 2015.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cai, D. and Lam, W. Graph transformer for graph-to-sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7464–7471, 2020.

Cho, K., van Merrienboer, B., et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

Corso, G., Cavalleri, L., et al. Principal neighbourhood aggregation for graph nets. *NeurIPS*, 2020.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., and Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

Devlin, J., Chang, M.-W., and others. Bert: Pre-training of deep bidirectional transformers for language understanding. *NACCL*, 2019.

Du, Y., Fu, T., Sun, J., and Liu, S. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*, 2022.

Ertl, P. An algorithm to identify functional groups in organic molecules. *Journal of cheminformatics*, 2017.

Fang, X., Liu, L., and others. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 2021.

Fang, Y., Zhang, Q., and others. Molecular contrastive learning with chemical element knowledge graph. *AAAI*, 2022.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic acids research*, 36(suppl_1):D919–D922, 2007.

Hamilton, L. W., Ying, R., and Leskovec, J. Inductive representation learning on large graphs. *NIPS*, 2017.

Han, X., Huang, Z., and others. Adaptive transfer learning on graph neural networks. In *KDD*, 2021.

Hao, Y., Dong, L., and others. Visualizing and understanding the effectiveness of bert. *EMNLP/IJCNLP*, 2019.

Hasanzadeh, A., Hajiramezanali, E., and others. Semi-implicit graph variational auto-encoders. *NeurIPS*, 2019.

Hassani, K. and Khasahmadi, A. Contrastive multi-view representation learning on graphs. In *ICML*, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3079209.

Hou, T., Wang, J., Zhang, W., and Xu, X. Adme evaluation in drug discovery. 6. can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *Journal of Chemical Information and Modeling*, 2007.

Hu*, W., Liu*, B., and others. Strategies for pre-training graph neural networks. In *ICLR*, 2020.

Hu, Z., Dong, Y., and others. Gpt-gnn: Generative pre-training of graph neural networks. In *KDD*, 2020.

Hussain, M. S., Zaki, M. J., and Subramanian, D. Edge-augmented graph transformers: Global self-attention is enough for graphs. *arXiv preprint arXiv:2108.03348*, 2021.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

Kipf, N. T. and Welling, M. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv:1611.07308*, 2016.

Li, P., Wang, J., and others. Pairwise half-graph discrimination: A simple graph-level self-supervised strategy for pre-training graph neural networks. In *IJCAI*, 2021a.

Li, P., Wang, J., and others. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *BIB*, 2021b.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. *arXiv preprint arXiv:2109.11730*, 2021c.

Liu, H., Sun, J., Guan, J., Zheng, J., and Zhou, S. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12): i221–i229, 2015.

Liu, S., Wang, H., and others. Pre-training molecular graph representation with 3d geometry. In *ICLR*, 2022.

Liu, Y., Pan, S., and others. Graph self-supervised learning: A survey. *arXiv:2103.00111*, 2021.

Lu, Y., Jiang, X., Fang, Y., and Shi, C. Learning to pre-train graph neural networks. In *AAAI*, 2021.

Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *Journal of medicinal chemistry*, 2011.

Mesquita, D., Souza, A. H., and Kaski, S. Rethinking pooling in graph neural networks. In *NeurIPS*, 2020.

Mialon, G., Chen, D., Selosse, M., and Mairal, J. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:2106.05667*, 2021.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021.

Pan, S., Hu, R., and others. Adversarially regularized graph autoencoder for graph embedding. In *IJCAI*, 2018.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Rong, Y., Bian, Y., and others. Self-supervised graph transformer on large-scale molecular data. *NeurIPS*, 2020.

Ryu, J. Y., Kim, H. U., and Lee, S. Y. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 115(18):E4304–E4311, 2018.

Scheiber, J., Chen, B., Milik, M., Sukuru, S. C. K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., et al. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *Journal of chemical information and modeling*, 49(2):308–317, 2009a.

Scheiber, J., Jenkins, J. L., Sukuru, S. C. K., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J., Urban, L., et al. Mapping adverse drug reactions in chemical space. *Journal of medicinal chemistry*, 52(9):3103–3107, 2009b.

Schütt, K., Kindermans, P., et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NIPS*, 2017.

Schwaller, P., Vaucher, A. C., Laino, T., and Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2 (1):015016, 2021.

Segler, M. H., Preuss, M., and Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

Stärk, H., Beaini, D., and others. 3d infomax improves gnns for molecular property prediction. *arXiv:2110.04126*, 2021.

Sun, F.-Y., Hoffman, J., Verma, V., and Tang, J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1lfF2NYvH.

Sun, M., Xing, J., and others. Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge. *KDD*, 2021.

Suresh, S., Li, P., and others. Adversarial graph augmentation to improve graph contrastive learning. In *NeurIPS*, 2021.

T., S., T., C., and others. Bootstrapped representation learning on graphs. In *ICLR Workshop*, 2021.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., and Aittokallio, T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

Tatonetti, N. P., Ye, P. P., Daneshjou, R., and Altman, R. B. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.

Vaswani, A., Shazeer, N., and others. Attention is all you need. *NIPS*, 2017.

Velickovic, P., Cucurull, G., and others. Graph attention networks. *ICLR*, 2018.

Velickovic, P., Fedus, W., and others. Deep graph infomax. *ICLR*, 2019.

Wang, C., Pan, S., and others. Mgae: Marginalized graph autoencoder for graph clustering. In *CIKM*, 2017.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., et al. Pubchem's bioassay database. *Nucleic acids research*, 40(D1):D400–D412, 2012.

Wang, Y., Wang, J., and others. Molclr: Molecular contrastive learning of representations via graph neural networks. *ArXiv*, abs/2102.10056, 2021.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906, 2008.

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

Xia, J., Lin, H., Xu, Y., Wu, L., Gao, Z., Li, S., and Li, S. Z. Towards robust graph neural networks against label noise, 2021. URL https://openreview.net/forum?id=H38f_9b90BO.

Xia, J., Wu, L., Chen, J., Hu, B., and Li, S. Z. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 1070–1079, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512156. URL https://doi.org/10.1145/3485447.3512156.

Xia, J., Wu, L., Wang, G., and Li, S. Z. Progcl: Rethinking hard negative mining in graph contrastive learning. In *International conference on machine learning*. PMLR, 2022b.

Xia, J., Zheng, J., Tan, C., Wang, G., and Li, S. Z. Towards effective and generalizable fine-tuning for pre-trained molecular graph models. *bioRxiv*, 2022c. doi: 10.1101/2022.02.03.479055.

Xia, T. and Ku, W. Geometric graph representation learning on protein structure prediction. In *KDD*, 2021.

Xie, Y., Xu, Z., and others. Self-supervised learning of graph neural networks: A unified review. *arXiv:2102.10757*, 2021.

Xu, D., C., W., and others. InfoGCL: Information-aware graph contrastive learning. In *NeurIPS*, 2021a.

Xu, K., Hu, W., and others. How powerful are graph neural networks? In *ICLR*, 2019.

Xu, M., Wang, H., and others. Self-supervised graph-level representation learning with local and global structure. In *ICML*, 2021b.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Yang, Z., Cui, Y., et al. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing. In *ACL: System Demonstrations*, 2020.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=OeWooOxFwDa.

You, Y., Chen, T., and others. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

You, Y., Chen, T., and others. Graph contrastive learning automated. *ICML*, 2021.

You, Y., Chen, T., Wang, Z., and Shen, Y. Bringing your own view: Graph contrastive learning without prefabricated data augmentations. In *WSDM*, 2022.

Zhang, H., Wu, Q., and others. From canonical correlation analysis to self-supervised graph neural networks. In *NeurIPS*, 2021a.

Zhang, J., Xiao, X., Huang, L.-K., Rong, Y., and Bian, Y. Fine-tuning graph neural networks via graph topology induced optimal transport. *arXiv preprint arXiv:2203.10453*, 2022.

Zhang, Z., Liu, Q., and others. Motif-based graph self-supervised learning for molecular property prediction. *NeurIPS*, 2021b.

Zhu, J., Xia, Y., Qin, T., gang Zhou, W., Li, H., and Liu, T.-Y. Dual-view molecule pre-training. *ArXiv*, abs/2106.10234, 2021.