Graph-based Molecular Representation Learning

Zhichun Guo\textsuperscript{1}, Kehan Guo\textsuperscript{1}, Bozhao Nan\textsuperscript{1}, Yijun Tian\textsuperscript{1}, Roshni G. Iyer\textsuperscript{2}, Yihong Ma\textsuperscript{1}, Olaf Wiest\textsuperscript{3}, Xiangliang Zhang\textsuperscript{4}, Wei Wang\textsuperscript{2}, Chuxu Zhang\textsuperscript{3}, Nitesh V. Chawla\textsuperscript{1}

\textsuperscript{1}University of Notre Dame
\textsuperscript{2}University of California, Los Angeles
\textsuperscript{3}Brandeis University

\{zguo5,kguo2,ban,yijun.tian,yma5,Olaf.G.Wiest.1,xzhang33,nchawla\}@nd.edu
\{roshnigiyer,weiwang\}@cs.ucla.edu, chuxuzhang@brandeis.edu

Abstract

Molecular representation learning (MRL) is a key step to build the connection between machine learning and chemical science. In particular, it encodes molecules as numerical vectors preserving the molecular structures and features, on top of which the downstream tasks (e.g., property prediction) can be performed. Recently, MRL has achieved considerable progress, especially in methods based on deep molecular graph learning. In this survey, we systematically review these graph-based molecular representation techniques, especially the methods incorporating chemical domain knowledge. Specifically, we first introduce the features of 2D and 3D molecular graphs. Then we summarize and categorize MRL methods into three groups based on their input. Furthermore, we discuss some typical chemical applications supported by MRL. To facilitate studies in this fast-developing area, we also list the benchmarks and commonly used datasets in the paper. Finally, we share our thoughts on future research directions.

1 Introduction

The interaction between machine learning and chemical science has received great attention from researchers in both areas. Remarkable progress has been made by applying machine learning in various chemical applications including molecular property prediction [Guo et al., 2020; Sun et al., 2021; Yang et al., 2021b; Liu et al., 2022c], reaction prediction [Jin et al., 2017; Do et al., 2019], molecular graph generation [Jin et al., 2018a; Jin et al., 2020b] and drug-drug interaction prediction [Lin et al., 2020]. Molecular representation learning (MRL) is an important step in bridging the gap between these two fields. MRL aims to utilize deep learning models to encode the input molecules as numerical vectors, which preserve relevant information about the molecules and serve as feature vectors for downstream chemical applications. While general representation learning models were earlier adapted to represent molecules, MRL algorithms have been recently designed to better incorporate chemical domain knowledge. However, it is non-trivial to have a seamless integration of domain knowledge into representation learning models. Given the tremendous effort in this rapidly-developing area, we are motivated to provide a systematic review of recent MRL methods, which are based on graph deep learning methods and integrate various types of chemical domain knowledge.

We focus on graph-based MRL for two reasons. First, molecules naturally lend themselves to graph representations, as they are essentially atoms and bonds interconnecting atoms. Compared with SMILES, a line-based representation (i.e., string) of molecules, molecular graphs provide richer information for MRL models to learn from. Accordingly, graph-based MRL models evolve much faster than sequence-based MRL models. Second, graph neural networks (GNN) have shown exceptional capacity and promising performance in handling graph structural data [Kipf and Welling, 2017; Hamilton et al., 2017; Zhang et al., 2019; Guo et al., 2023], certainly including those applied on molecular graphs [Gilmer et al., 2017; Hu et al., 2020; You et al., 2020]. It is thus urgent to summarize the effort of leveraging GNN on molecular graphs with domain knowledge and open this topic with more discussion.

This survey paper will be a contribution to both fields of machine learning and chemistry. A variety of molecule-centered problems can be formulated as predictive or generative tasks, e.g., molecule property prediction, reaction prediction, and molecule generation. The machine learning-enabled solutions for these problems have a common ground in learning high-quality representations for molecules. However, researchers in chemistry are overwhelmed by the big group of MRL models to choose from, not to mention that new models are rapidly presented. This survey provides an up-to-date overview of MRL models regarding the input graph to the models and the feasible downstream applications. Researchers in chemistry can easily find out the MRL models that match their application needs. For researchers in machine learning, a lack of understanding of the chemical domain knowledge is the barrier to addressing the representation learning for molecules. Treating molecular graphs as regular attributed graphs would overlook the special substructure patterns of molecules, such as motifs and functional groups. This survey summarizes the existing strategies for introducing chemistry-related domain knowledge into representation learning and will inspire researchers in machine learning to design more effective MRL models.
We organize the review in the following structure. We first introduce the input expression of molecules by 2D and 3D graphs. Following these categories of input, we summarize representative MRL algorithms regarding the usage of domain knowledge, the learning strategies, the application tasks, and code links if available. We then look into each application task and introduce in detail the usable MRL models and benchmark datasets. In the end, we open the discussion for future research directions and conclude.

2 Expression of Molecules

When machine learning was first introduced for molecular analysis, hand-crafted features based on predetermined fingerprint extraction rules were used to identify and represent significant information about molecules [Ahneman et al., 2018].

However, this feature engineering process can be time-consuming, expert-dependent, and may not always provide the best results. To address these challenges, deep learning models, known for representation learning, have been developed to learn important molecular features automatically. Molecule input to deep learning models can be presented in two kinds of expressions: molecular sequences and graphs.

The expression of molecular sequence, such as simplified molecular-input line-entry system (SMILES) [Weininger et al., 1989] and SELF-referencing Embedded Strings (SELFIES) [Krenn et al., 2020], may separate two connected atoms at two distant positions and lead to inferior representation.

In contrast, the expression by graph naturally incorporates additional information in nodes (atoms) and edges (bonds), which can be easily leveraged by the rich suite of graph-based models (e.g., graph neural networks). Therefore, MLR on molecular graphs is becoming commonly used and will be the focus of this survey.

In this section, we provide clarification on the distinction between 2D molecular graphs and 3D molecular graph representations, as shown in Figure 1. We analyze the characteristics of each representation and explore their applications and limitations when used in deep learning models.

| Attribute       | Details                                                                 |
|-----------------|-------------------------------------------------------------------------|
| Node            | Atom type: 118, Chirality tag: unspecified, tetrahedral cw, tetrahedral ccw, other, Hybridization: sp, sp², sp³-d, or sp³d², Atomaticity: 0 or 1 (atomic atom or not) |
| Edge            | Bond type: single, double, triple, aromatic, Ring: 0 or 1 (bond is in a ring or not), Bond direction: -, endupright, enddownright, Stereochemistry: -, any, Z, E, cis, trans, @ |

Table 1: Details of node and edge features in molecular graphs.

2.1 2D Molecular Graphs

A graph is typically composed of nodes connected by edges. Similarly, in a molecule, atoms can be seen as nodes and bonds as edges interconnecting them. Thus, each molecule has a natural graph structure. This renders molecular graphs to be the most feasible input for deep learning models and leads to their extensive use. The most common form of molecular graphs is described by three matrices: the node feature matrix, edge feature matrix, and adjacency matrix. Molecules are usually stored as SMILES for convenience and then converted to molecular graphs for computation using specific tools, such as RDKit [Landrum, 2020]. The commonly used features of nodes and edges are listed in Table 1, including mandatory features such as atom and bond types, and other optional features can be added as in need of different tasks [Tang et al., 2020; Saebi et al., 2023]. Among these features, the atom’s chirality tag cannot be learned from the common 2D molecular graph representation without 3D geometric information, while all other features are learnable from both 2D and 3D structures. Each bond is considered as a bidirectional edge, i.e., a bond between atoms A and B is given as two edges in the adjacency list: one from A to B, and the other from B to A. With the description matrices, 2D molecular graphs can be treated as homogeneous [Gilmer et al., 2017; Guo et al., 2021; Coley et al., 2019] or heterogeneous networks [Shui and Karypis, 2020] for learning molecular representations via leveraging graph neural networks (GNN).

Despite that GNN can be conveniently used on 2D molecular graphs, the learned representation neglects the spatial direction and torsion between atoms in molecules. This is mainly due to the limitation of 2D graph structure, which only presents the type of bond connecting two atoms, and has no information like torsion angle, bond length, and stereoisomerism. The missing information is important to catch the subtle difference that may cause a significant discrepancy in chemical problems. For instance, reactants with the same 2D graph structures can have different products, because the reactants may differ on torsion angle and bond length, which are not reflected in their 2D graphs.

2.2 3D Molecular Graphs

The 3D molecular graphs provide the missing geometric information by explicitly encoding the spatial structure of molecules. The 3D graphs present the atomic structure as a set of atoms along with their 3D coordinates, which in-
includes more spatial information about atoms. As a result, this representation format has received increasing attention in MRL [Liu et al., 2022c]. The key difference between 2D and 3D molecular graphs lies in the determination of edges between atoms. In 2D molecular graphs, edges (bonds) are pre-determined, indicating the existence of a connection or not. The soft edges of atomic interactions in 3D graphs can be determined by the distance and also the angle between two atoms using their coordinates.

To incorporate more complicated spatial relationships, spherical graph neural network [Liu et al., 2022c] is designed to learn molecule structure from 3D graphs.

3 Methodologies of MRL

In this section, we summarize MRL methods into three categories based on the types of input molecules: 2D-based, 3D-based, and knowledge graph-based MRL methods. We introduce the encoding method for each category and point out recent representative methodologies (summarized in Table 2).

3.1 2D-based MRL Methods

2D molecular graphs are the most widely used inputs for graph-based MRL. Here, we introduce the general graph neural networks to learn molecular representations with 2D graphs. Then, following several chemical substructure definition clarifications, we will chart the path from the general representation learning methods to the representative methods incorporating molecular structures and chemistry-related domain knowledge.

Encoding methods. Formally, each molecule generally is considered as an undirected graph \( G = (V, E, X) \) where node features \( x_v \in X \) for \( v \in V \) and edge features \( e_{uv} \in E \) for \( (u, v) \in E \) [Brockschmidt, 2020]. Here, nodes represent atoms and edges represent bonds. Generally, graph-based learning methods can fit into Message Passing Neural Networks (MPNN) [Gilmer et al., 2017] scheme. Therefore, we take MPNN as an example to illustrate the learning process. The forward pass consists of three operations: message passing, node update, and readout. During the message passing phase, node features are updated iteratively according to their neighbors in the graph for \( T \) times. By initializing the embedding of each node \( v \) as \( h_0^v = x_v \), node hidden states \( h_t^v \) at step \( t + 1 \) are obtained based on messages \( m_{uv}^{t+1} \), which are represented as:

\[
m_{uv}^{t+1} = \sum_{u \in N(v)} M_t(h_u^v, h_v^u, e_{uv})
\]

and

\[
h_{v}^{t+1} = U_t(h_v^t, m_{uv}^{t+1}),
\]

where \( M_t \) denotes the message function, \( U_t \) is the node update function, and \( N(v) \) is the set of node \( v \)'s neighbors in the graph. After updating the node features \( T \) times, the readout function \( R \) computes the whole graph embedding vector by \( y = R(h_v^T \mid v \in V) \). Note that \( R \) is invariant to the order of nodes so that the framework can be invariant to graph isomorphism. \( y \) is the representation for the molecule and can be passed to downstream tasks. All functions \( M_t, U_t, \) and \( R \) are neural networks with learnable weights to update during the training process.

Besides MPNN, different variants of graph neural networks like GCN [Kipf and Welling, 2017], GIN [Xu et al., 2019], GAT [Velickovic et al., 2018], GGNN [Li et al., 2016], GraphSage [Hamilton et al., 2017], HetGNN [Zhang et al., 2019] and feature-wise GNN-Film [Brockschmidt, 2020] can also be used directly to learn molecular representations. These methods are widely utilized as the base encoder for molecular representation learning in various downstream tasks, such as reaction prediction [Coley et al., 2019], property prediction [Brockschmidt, 2020] and drug discovery [Jin et al., 2020c]. Hu et al. [Hu et al., 2020] conduct a comparative study on GNNs in property prediction and find that GIN usually achieves the best results. While these models are powerful in learning graph structures, chemical traits, and other chemical domain knowledge are largely neglected.

Chemical Substructures. Molecular graphs have substructures that are relevant to certain molecular properties or represent molecular generation constraints. These substructures are not just subgraphs, and convey special domain knowledge. They are clarified and distinguished next. Chemical substructures [Jin et al., 2018a] could be clusters of atoms, e.g., rings. Motifs [Rong et al., 2020; Sun et al., 2021] are recurrent sub-graphs among the input graphs. The functional group is an important component of motifs and encodes rich domain knowledge of molecules. This type of motif can be detected by RDKit [Landrum, 2020]. A Rationale [Jin et al., 2020c; Liu et al., 2022a] is a sub-graph that has a particular molecular property. Scaffolds [Maziarz et al., 2022] are predefined chemical sub-graphs. Structures sharing the same scaffold can always be considered to be generated following the same synthetic pathway.

Learning Strategies. The weights in MRL encoder can be trained in an end-to-end fashion by attaching the encoder with a downstream task, e.g., the representation after encoding is sent to make property prediction by a fully connected layer. The training thus takes a supervised manner. To enhance the training process, molecular substructure-related or chemical domain knowledge can be utilized. JT-VAE [Jin et al., 2018a], RationaleRL [Jin et al., 2020c], and HierVAE [Jin et al., 2020b] take advantages of chemical substructures, rationales, and motifs respectively for the molecular generation. Yang et. al [Yang et al., 2021b] proposed PhysChem. This method is composed of a physicist network to learn the molecular conformation and a chemist network to learn the molecular properties. It shows good performance on property prediction benchmarks by fusing both chemical and physics information. Wang et. al [Wang et al., 2021] involved task information and proposed a property-aware embedding method for molecular property prediction.

Although the learned representation may perform well for the specific downstream task, it is not generally usable for other tasks. In addition, supervised training requires a sufficient set of annotated training samples, which are often difficult to acquire. Recent research has seen a foray into self-supervised learning strategies [Jin et al., 2020a] that propose reconstruction tasks for pre-training. PreGNN [Hu et al., 2020] uses two self-supervised strategies (context prediction and node/edge attribute masking) to pre-train GNN. GROVER [Rong et al., 2020] involves molecular-specific self-supervised methods: contextual property prediction and graph-level motif prediction. Zhang et. al [Zhang et al., 2021] also designed a motif-based graph self-supervised strategy, which predicted the motif’s topology and label during the motif tree generation process.
Contrastive learning is another common self-supervised learning technique, where augmented graphs are biased to keep close to the anchor graph (positive pair) and away from other graphs (negative pairs). It can help graph encoder models to produce graph representations with better generalizability, transferability, and robustness. The general graph augmentation methods (node dropping, edge perturbation, attribute masking and subgraph) were proposed by You et al. [You et al., 2020], which could be applied to molecular datasets as well. InfoGraph [Sun et al., 2020] trains the model by maximizing the mutual information between the representations of the entire graph and substructures of different granularity. Unlike general contrastive learning strategies, the following models incorporate chemical domain knowledge. MoCL [Sun et al., 2021] has two proposed molecular graph augmentation methods: one is replacing a valid substructure with a similar physical or a chemical property-related substructure. The other one is changing a few general carbon atoms. Wang et al. [Wang et al., 2022a] were inspired by the relation of equivalence between reactants and products in a chemical reaction. They proposed MoIR to preserve the equivalence relation in the embedding space. It forces the sum of reactant embeddings and the sum of product embeddings to be equal. The reactant and the product from different reactions could be a negative pair for this contrastive learning.

### 3.2 3D-based MRL Methods
Molecular spatial information, especially geometric information, attracts wide attention and has been increasingly involved in the MRL in recent years, especially when the model needs to learn the physical and chemical properties of molecules associated with the 3D positions of atoms.

**Encoding methods.** 3D molecule is a dynamic structure since atoms are in continual motion in 3D space. The local minima on the potential energy surface are called conformer, or conformation. In nature, a molecule contains multiple low-energy conformers exhibiting different chemical properties. Therefore, the 3D molecular graph implicitly encodes spatial positions of atoms to learn better representation. In general, the 3D conformer of a molecule can be denoted as $G^{3D} = (V, C)$ where $V$ is the node set, and $C$ is the 3D-coordinate matrix. Various encoding methods are explored to enhance GNNs with a proper 3D-coordinate matrix. GeomCL [Li et al., 2022] and GRAPHMVP [Liu et al., 2022b] utilize conformers generated from RDkit through the stochastic optimization algorithm using the Merck Molecular Force Field (MMFF) to encode spatial information such as angles, interatomic distances, torsion, etc. SphereNet [Liu et al., 2022c]...

| Input | Algorithm | Encoder | Pretrain | Domain Knowledge | Tasks | Venue | Code Link |
|-------|-----------|---------|----------|------------------|-------|-------|-----------|
| 2D    | MPNN[1]   | MPNN    | /        | /                | PP    | ICML’17| /                     |
|       | Pre-GNN[2]| GIN     | RT       | /                | PP    | ICLR’20| https://github.com/snap-stanford/pretrain-gnn/ |
|       | InfoGraph[3]| GNN   | CL       | /                | PP    | ICLR’20| https://github.com/fanyun-sun/InfoGraph |
|       | GNN-FILM[4]| MPNN   | /        | /                | PP    | ICLR’20| https://github.com/Microsoft/tf-gnn-samples |
|       | GROVER[5] | GAT     | RT       | Motif            | PP    | NeurIPS20 | https://github.com/enzimetlab/grover |
|       | GraphCL[6] | GNN    | CL       | /                | PP    | NeurIPS20| https://github.com/tiankai-lab/GraphGL |
|       | MoCL[7]   | GIN     | CL       | Motif&General carbon | PP    | KDD’21 | https://github.com/minghui2098/MoCL |
|       | MCSSS[8]  | GIN     | RT       | Motif            | PP    | NeurIPS21| https://github.com/raizin/RISMG |
|       | PhysChem[9]| MPNN   | /        | QtyNet and ChemNet | PP    | NeurIPS21| /                     |
|       | GERA[10]  | GNN     | /        | Rationale        | PP    | KDD’22 | https://github.com/liusongcode/GERA |
|       | MoleOOD[11]| SAGE   | /        | Scaffold          | PP    | NeurIPS22| https://github.com/yassinsiana/moleOOD |
|       | JT-VAE[12]| MPNN   | /        | Chemical substructure | MG   | ICLR’18| https://github.com/seganp-jit/icom18-jnsm |
|       | VTNNS[13]| MPNN    | /        | Chemical substructure | MG   | ICLR’18| https://github.com/seganp-jit/icom18-jnsm |
|       | GraphAF[14]| R-GCN  | /        | Valency constraint | MG   | ICLR’20 | https://github.com/DeepGraphLearning/GraphAF |
|       | RationalE[15]| MPNN   | /        | Rationale        | MG   | ICLR’20 | https://github.com/mouten/rationale |
|       | HieReA[16]| MPN    | /        | Motif            | MG   | ICLR’20 | https://github.com/mouten/rationale |
|       | MoLeR[17] | GNN     | /        | Motif            | MG   | ICLR’22 | /                     |
|       | WLDN[18]  | WLN     | /        | /                | RP    | NeurIPS17| https://github.com/seganp-jit/rips17-rongen |
|       | MolIR[19] | GNN     | CL       | Reaction equivalence | PP    | ICLR’22 | https://github.com/huang1175/MolIR |
| 3D    | DimeNet[20]| MPNN   | /        | Spatial           | PP    | ICLR’19 | https://github.com/KLer/pajilab/dimenes |
|       | SphereNet[21]| MPN   | /        | Spatial           | PP    | ICLR’22 | https://github.com/divelab/DSC |
|       | ConVAE[22] | GNN    | /        | Spatial           | PP    | ICLR’22 | https://github.com/klasi/ChemVAE-ICB11 |
|       | GRAPHMVP[23]| GNN   | CL       | Spatial           | PP    | ICLR’22 | https://github.com/zhao22/graphMVP |
|       | 3D-Informax[24]| MPNN  | CL       | Spatial           | PP    | ICLR’22 | https://github.com/randolftk/3D-Informax |
|       | UnifiedPMI[25]| GBLOCKS| SPatial   | /                | PP    | KDD’22 | https://github.com/ivaleo/moleiod/pdppretrain |
|       | GeomGCL[26]| MPNN  | CL       | Spatial           | PP    | AAA1’22| https://github.com/zhc-fangyin/GeomGCL |
| KG    | KGNN[27]  | GNN     | MPNN      | Drug-drug interaction | DDI  | IJCAI’20| https://github.com/seganp-jit/RKB |
|       | KCT[28]   | GNN     | MPNN      | Chemical element  | DDI  | AAA1’22| https://github.com/ZF Fangyi/KCT |

| [1] Gilmer et al., 2017; [2] Hu et al., 2020; [3] Sun et al., 2020; [4] Brockschmidt, 2020; [5] Rong et al., 2020; [6] You et al., 2020; [7] Sun et al., 2021;[8] Zhang et al., 2021; [9] Yang et al., 2021b; [10] Liu et al., 2022a; [11] Yang et al., 2022; [12] Jia et al., 2018a; [13] Jia et al., 2018b; [14] Shi et al., 2020; [15] Jia et al., 2020c; [16] Jia et al., 2020b; [17] Maziarz et al., 2022; [18] Jia et al., 2017; [19] Wang et al., 2022a; [20] Klicpera et al., 2019; [21] Liu et al., 2022c; [22] Xu et al., 2021; [23] Liu et al., 2022b; [24] Skark et al., 2022; [25] Zhu et al., 2022; [26] Li et al., 2022; [27] Lin et al., 2020; [28] Fang et al., 2022; |
employs sphere coordinates and designs a spherical message passing as a powerful scheme for 3D molecular learning. Follow this work, ComENet [Wang et al., 2022b] is a message-passing paradigm on 3D molecular graphs with better completeness (bijectivity) by leveraging rotation angles. To circumvent the challenge that true 3D coordinates of molecules are difficult to calculate and sometimes non-deterministic, DimeNet [Klicpera et al., 2019], GemNet [Klicpera et al., 2021a] and Directional MPNN [Klicpera et al., 2021b] generate synthetic coordinates in molecules through computing molecular distance bound and corresponding angles between atom pairs, where directional message passing networks are applied to learn the enhanced representation.

**Learning Strategies.** In addition to supervised learning, recent MRL studies propose self-supervised learning techniques by using both 2D and 3D molecular graphs. UnifiedPML [Zhu et al., 2022] uses three pre-training reconstruction tasks: reconstruction of masked atom and coordinates, 3D conformation generation based on 2D graph, and 2D graph generation based on 3D conformation. For contrastive learning, molecular 2D and 3D graph representations are naturally two augmented views of molecules. Using this characteristic, GeomGCL [Li et al., 2022], GraphMVP [Liu et al., 2022b], and 3D-Informax [Stärk et al., 2022] were proposed to train the molecular representations by keeping the consistency between 2D and 3D graph information. GeomGCL utilizes 2D geometric information, while GraphMVP and 3D-Informax use 2D topological information. Different from the other two methods, 3D-Informax utilizes multiple 3D conformers instead of one.

### 3.3 Knowledge Graph-based MRL Methods

The knowledge graph-based methods are proposed to involve molecular-structure-invariant but rich external knowledge in the model. KCL [Fang et al., 2022] is a molecular augmentation method for contrastive learning with an external knowledge graph, which is formed by triples in the form of (chemical element, relation, attribute), such as (Gas, isStateOf, Cl). Then a new node “Gas” connecting with “Cl” atom will be generated to the original 2D molecular graph. After augmentation, the model will be trained by maximizing the agreement between two views of molecular graphs with a contrastive loss function. In contrast to KCL and the above MRL methods, which take atoms as nodes and bonds as edges forming a graph, KGNN [Lin et al., 2020] and MDNN [Lyu et al., 2021] explore the knowledge graph consisting of molecules as nodes and connection relationship between molecules as edges. In this case, molecular representations are learned from the knowledge graph structures instead of molecular structures.

### 4 Application Tasks of MRL

In this section, we discuss four real applications of MRL, and present the details of representative work in Table 3.

#### 4.1 Property Prediction

Molecular property prediction plays a fundamental role in drug discovery to identify potential drug candidates with target properties. Generally, this task consists of two phases: a molecular encoder to generate a fixed-length molecular representation and a predictor. A predictor is utilized to predict whether the molecule has the target property or predict the reaction of molecules to the target property based on learned molecular representation. Property prediction results can reflect the quality of learned molecular representation directly. General graph learning papers [Hu et al., 2020; Gilmer et al., 2017; Brockschmidt, 2020; You et al., 2020] thus take property prediction tasks to evaluate the performance of their algorithms. Compared with these general methods, the pre-training methods (e.g., GROVER, MoCL, and MGSSL) involving molecular substructure-related designs are more appropriate and can mostly achieve better prediction performance. Some recent work proposes supervised learning models specifically for property prediction. GREA [Liu et al., 2022a] defines the complementary subgraph of rationale in each molecular graph as an environmental subgraph of that molecule, which has no relationship with the target property. Yang et al. [Yang et al., 2022] proposed MoleOOD to guide the encoder to focus on the environment-invariant substructures to improve its generalization ability. Among the above methods, UnifiedPML [Zhu et al., 2022] achieves SOTA performance by leveraging the unified view of both 2D and 3D molecular graphs. Besides, the insufficient available molecular dataset is a common problem ex-

| Methods | Reference | Evaluation Metrics |
|---------|-----------|--------------------|
| Molecule Generation | Validity | Diversity | Others |
| JT-VAE | [Jin et al., 2018a] | ✓ | ✓ | Reconstruction |
| GraphAF | [Shi et al., 2020] | ✓ | ✓ | Reconstruction |
| RationalRL | [Jin et al., 2020c] | ✓ | ✓ | Success |
| HierVAE | [Jin et al., 2020b] | ✓ | ✓ | Reconstruction |
| MolLe* | [Maziari et al., 2022] | ✓ | ✓ | FCD |
| ContVAE | [Xu et al., 2021] | ✓ | ✓ | COV & MAF |
| UnifiedPML* | [Zhu et al., 2022] | ✓ | ✓ | COV & MAF |

| Reaction Prediction | Coverage & Accuracy | MRR & Hits |
|---------------------|---------------------|------------|
| WLDN++ | [Coley et al., 2019] | ✓ | ✓ | ✓ |
| MolR* | [Wang et al., 2022a] | ✓ | ✓ | ✓ |

| Drug-drug Interactions | AUC | P-AUC | ACC | F1 |
|------------------------|-----|-------|-----|---|
| AntiSemiGAE | [Ma et al., 2018] | ✓ | ✓ | ✓ |
| KGNN | [Lin et al., 2020] | ✓ | ✓ | ✓ |
| MDNN* | [Lyu et al., 2021] | ✓ | ✓ | ✓ |

Table 3: Applications for MRL with representative methods and evaluation metrics for each method. The methods marked with “*” denote SOTA for each application. AUC indicates ROC-AUC.
Table 4: Datasets used in MRL study.

| Dataset       | Category                  | #Train | #Dev | #Test | Reference              | Data Link                  |
|---------------|---------------------------|--------|------|-------|------------------------|---------------------------|
| ZINC15        | Structure Pretraining     | 40,000 | 4,112| 4,112 | [Sterling and Irwin, 2015] | https://zinc15 docking.org |
| PubChem       | Structure Pretraining     | 40,000 | 4,112| 4,112 | [Kim et al., 2019]     | https://pubchem.ncbi.nlm.nih.gov |
| ChEMBL        | Structure Pretraining     | 40,000 | 4,112| 4,112 | [Gaulton et al., 2017]  | https://www.ebi.ac.uk/chembl/ |
| QM9           | Property prediction       | 107,108| 13,388| 13,388| [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| ESOL          | Property prediction       | 902    | 112  | 112   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| FreeSolv      | Property prediction       | 513    | 64   | 64    | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| Lipophilicity | Property prediction       | 3,360  | 420  | 420   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| MUV           | Property prediction       | 74,470 | 9,308| 9,308 | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| HIV           | Property prediction       | 32,901 | 4,112| 4,112 | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| PDBbind       | Property prediction       | 9,526  | 1,190| 1,190 | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| BACE          | Property prediction       | 1,210  | 151  | 151   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| BBBP          | Property prediction       | 1,631  | 203  | 203   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| Tox21         | Property prediction       | 6,264  | 783  | 783   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| ToxCast       | Property prediction       | 6,860  | 857  | 857   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| SIDER         | Property prediction       | 1,141  | 142  | 142   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| ClinTox       | Property prediction       | 1,182  | 147  | 147   | [Wu et al., 2018]      | https://moleculenet.org/datasets-1 |
| USPTO_MIT     | Reaction Prediction       | 400,000| 40,000| 40,000| [Jin et al., 2017]    | https://github.com/wengong jin/nips17 rexgen |
| USPTO-15K     | Reaction Prediction       | 10500  | 1500 | 1500  | [Coley et al., 2017]   | https://github.com/commongen/chempredict_pm |
| USPTO-full    | Reaction Prediction       | 760,000| 95,000| 95,000| [Lowe, 2012]          | https://github.com/danb80/patent reaction extraction |
| ZINC-250k     | Molecular Generation     | 200,000| 25,000| 25,000| [Kusner et al., 2017]  | https://github.com/dastan/primemolaeS |
| DrugBank      | Drug-drug interaction     | 489,910| 61,238| 61,238| [Lin et al., 2020]     | https://github.com/xzenglab/KGNN |
| KEGG-drug     | Drug-drug interaction     | 45,586 | 5,698| 5,698 | [Lin et al., 2020]     | https://github.com/xzenglab/KGNN/treen master |

existing in chemistry. Guo et al. [Guo et al., 2021] and Wang et al. [Wang et al., 2021] proposed meta-learning methods to deal with this low-data problem on property prediction.

The property prediction task is mainly conducted as classification and regression. The classification task is to predict a discrete class label, for which the Area under the ROC curve (ROC-AUC) and accuracy (ACC) are selected evaluation metrics [Sun et al., 2020; Rong et al., 2020]. The regression task is to predict a continuous quantity for each molecule, where the mean absolute error (MAE) and Root Mean Square Error (RMSE) are commonly used evaluation metrics to provide the estimate of the accuracy for the target [Rong et al., 2020; Zhu et al., 2022].

4.2 Molecular Generation

The key challenge of drug discovery is to find target molecules with the target properties, which heavily relies on domain experts. The molecular generation is to automate this process. Two steps are necessary to complete this task: one is designing an encoder to represent molecules in a continuous manner, which is beneficial to optimize and predict property; the other is proposing a decoder to map the optimized space to a molecular graph with the optimized property. Atom-by-atom generations may generate atypical chemical substructures such as partial rings. To avoid invalid states [Jin et al., 2018a], most studies generate graphs fragment by fragment instead of node by node. JT-VAE [Jin et al., 2018a] and VJTN [Jin et al., 2018b] decompose the molecular graph into the junction tree first, based on substructures in the graph. Then they encode the junction tree using a neural network. Next, they reconstruct the junction tree and assemble nodes in the tree back to the original molecular graph. These methods are highly complex and frequently fail for substructures involving more than 10 atoms. Different from the above substructures, motifs have much larger and more flexible substructures. Therefore, to deal with the problems existing in the previous methods, Jin et al. [Jin et al., 2020b] proposed HierVAE, which relies on motifs instead of substructures and generates molecular graphs hierarchically based on motifs. Following this direction, MoLeR [Maziarz et al., 2022] was proposed to generate molecules by combining atom-by-atom and motif-by-motif generation. What’s more, it supports the scaffolds as an initial seed for molecular generation, which outperforms previous methods on scaffold-constrained tasks. Besides substructures and motifs, RationaleRL [Jin et al., 2020c] utilizes Monte Carlo Tree Search to extract the subgraphs from the molecules and construct rationale vocabulary for each property with their predictive models. It composes molecules from the sampled rationales to preserve the properties of interest with a variational auto-encoder. In contrast to the above encoder-decoder framework, GraphAF [Shi et al., 2020] is a flow-based auto-regressive model to generate the molecular graph in a sequential process, which allows it to leverage chemical valency constraints in each generation step and also enjoys efficient parallel computation in the training process. ConfVAE [Xu et al., 2021] and UnifiedPML [Zhu et al., 2022] were proposed for molecular 3D conformation generation tasks.

For molecular generation tasks, validity is the percentage of the chemically valid generated molecules [Jin et al., 2018a]. Diversity measures the diversity of the generated positive compounds by computing their distance in chemical space. Reconstruction accuracy is utilized to evaluate how often the model can reconstruct a given molecule from the latent embedding [Jin et al., 2020b]. Frechet ChemNet Distance (FCD) is utilized to measure how much the sampled molecules resemble the training molecules [Maziarz et al., 2022]. For property-constraints molecular generation tasks, the model also needs to measure the fraction of generated molecules that match the target property. For molecular conformation generation tasks, coverage score (COV) and matching score (MAT) are usually utilized for evaluation [Xu et al., 2021; Zhu et al., 2022].
4.3 Reaction Prediction

Reaction prediction and retrosynthesis prediction are fundamental problems in organic chemistry. Reaction prediction means using reactants to predict reaction products. The process of retrosynthesis prediction is the opposite of reaction prediction. When taking SMILES as input, the reaction prediction task is analogous to a translation task. When taking molecular graphs as input, there are two steps to do both for reaction prediction and retrosynthesis prediction. Like WLDN [Jin et al., 2017] and WLDN++ [Coley et al., 2019], the model needs to predict the reaction center first and then predict the potential products which is the major product. These tasks will be evaluated by coverage (whether the candidates cover the correct product) and accuracy (whether the model can select the correct product). Different from previous work, MolR [Wang et al., 2022a] formulates the task of reaction prediction as a ranking problem. All the products in the test set are put in the candidate pool. MolR ranks these candidate products based on the embedding learned from given reactant sets, using mean reciprocal rank (MRR) and hits ratio (Hits) as the evaluation metrics.

4.4 Drug-drug Interactions

Detecting drug-drug interaction (DDI) is an important task that can help clinicians make effective decisions and schedule appropriate therapy programs. Accurate DDI can not only help medication recommendations but also effectively identify potential adverse effects, which is critical for patients and society. AttSemiGAE [Ma et al., 2018] predicts DDI by measuring drug similarity with multiple types of drug features. SafeDrug [Yang et al., 2021a] designs global and local two modules to fully encode the connectivity and functionality of drug molecules to predict DDI. MoleRec [Yang et al., 2023] proposes a molecular substructure-aware representation learning strategy for DDI. Both KGNN [Lin et al., 2020] and MDNN [Lyu et al., 2021] build the drug knowledge graph to improve the accuracy of DDI. DDI prediction is evaluated according to Accuracy (ACC), ROC-AUC (AUC), P-AUC (area under the precision-recall-curve), and F1 scores.

5 Molecular Datasets and Benchmarks

We summarize representative molecular representation learning algorithms in Table 2. To conveniently access the empirical results, each paper is attached with code links if available. Encoding algorithms, pre-training methods, and the utilized domain knowledge are also listed. Here, pre-training methods specify contrastive learning and reconstruction tasks we discussed in Section 3. We also present the representative methods for each application and their corresponding evaluation metrics in Table 3. The SOTA for each application is also labeled. In addition, we summarize commonly used datasets for different chemical tasks in Table 4.

6 Future Directions

Graph-based methods for MRL develop fast. Although MRL has achieved satisfactory results in various applications, there are still some challenges that remain to be solved. We list several future directions for reference.

6.1 Graph-based MRL with Spatial Learning

The 3D geometric information attracts great attention recently in graph-based MRL. There are several ways to encode 3D information. One is an equivariant graph neural network, like SE(3)-transformers [Fuchs et al., 2020]. Another category of methods takes relative 3D information as input, like the directional message passing methods [Klicpera et al., 2019; Klicpera et al., 2021a] introduced in Section 3, which include distances between atoms and angles between bonds as features to learn geometric information. SphereNet [Liu et al., 2022c] leverages spherical message passing to learn 3D molecular representation. Nevertheless, how different geometries contribute to molecular representation learning still lacks rigorous justification. There is no established standard spatial information learning method for now. It should be a promising future research direction for MRL.

6.2 Graph-based MRL with Explainability

The explainability is always a challenge for MRL. To break down the gap between machine learning and chemical science, a well-designed MRL model to produce competitive prediction or generation results on chemical tasks is important but not the end of MRL research. Which molecular features play the most important part in MRL? How can MRL be helpful on explaining the process of reaction? How can MRL support the transparent generation of new drugs? The answers to these questions will facilitate the discovery and innovation in chemical science and engineering, as well as improving the trustworthiness of machine learning methods. AttSemiGAE [Ma et al., 2018], EE [Gao et al., 2018] and GCNN [Henderson et al., 2021] own initial strategies to improve their model’s explainability. However, explainable MRL remains a challenging research problem.

6.3 Graph-based MRL with Insufficient Data

Reliable data collection and annotation are time-consuming and expensive via experiments in the laboratory. As a result, data scarcity is a common problem in chemistry, and highly hinders the development of MLR. Self-supervised and meta-learning have been considered promising solutions in recent years. Guo et al. [Guo et al., 2021] and Wang et al. [Wang et al., 2021] proposed meta-learning algorithms to deal with few-shot molecule problems, which appeals to some following work. While only specific application tasks have been investigated, novel MRL algorithms should be further developed to deal with insufficient data problems.

7 Conclusion

Molecular representation learning builds a strong and vital connection between machine learning and chemical science. In this work, we introduce the problem of graph-based MRL and provide a comprehensive overview of the recent progresses on this research topic. To facilitate reproducible research, we take the first step to summarize and release the representative molecular representation learning benchmarks and commonly used datasets for the research community. This survey paper will be a useful resource for researchers in both chemistry and machine learning to advance the study of MRL and other molecular application tasks.
Acknowledgements
This work was supported by National Science Foundation under the NSF Center for Computer Assisted Synthesis (C-CAS), grant number CHE-2202693.

References
[Ahneman et al., 2018] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in c–n cross-coupling using machine learning. Science, 2018.

[Brockschmidt, 2020] Marc Brockschmidt. Gnn-film: Graph neural networks with feature-wise linear modulation. In ICLR, 2020.

[Coley et al., 2017] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. ACS central science, 2017.

[Coley et al., 2019] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. Chemical science, 2019.

[Do et al., 2019] Kien Do, Truyen Tran, and Svyeth Venkatesh. Graph transformation policy network for chemical reaction prediction. In KDD, 2019.

[Fang et al., 2022] Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhan, Shumin Deng, Wen Zhang, Ming Qin, Zuo Chen, Xiaohui Fan, and Huijuan Chen. Molecular contrastive learning with chemical element knowledge graph. In AAAI, 2022.

[Fuchs et al., 2020] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. NeurIPS, 2020.

[Gao et al., 2018] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In IJCAI, 2018.

[Gaulton et al., 2017] Anna Gaulton, Anne Hersey, Michal Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowa, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. Nucleic acids research, 2017.

[Gilmer et al., 2017] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In ICML, 2017.

[Guo et al., 2020] Zhichun Guo, Wenhai Yu, Chuxu Zhang, Meng Jiang, and Nitesh V Chawla. Graseq: graph and sequence fusion learning for molecular property prediction. In CIKM, 2020.

[Guo et al., 2021] Zhichun Guo, Chuxu Zhang, Wenhai Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. Few-shot graph learning for molecular property prediction. In WWW, 2021.

[Guo et al., 2023] Zhichun Guo, Chunhui Zhang, Yujie Fan, Yijun Tian, Chuxu Zhang, and Nitesh Chawla. Boosting graph neural networks via adaptive knowledge distillation. In AAAI, 2023.

[Hamilton et al., 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. NeurIPS, 2017.

[Henderson et al., 2021] Ryan Henderson, Djork-Arné Clevert, and Floriane Montanari. Improving graph neural network explainability with orthonormalization and induced sparsity. In ICML, 2021.

[Hu et al., 2020] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In ICLR, 2020.

[Jin et al., 2017] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-leman network. NeurIPS, 2017.

[Jin et al., 2018a] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In ICML, 2018.

[Jin et al., 2018b] Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. Learning multimodal graph-to-graph translation for molecule optimization. In ICLR, 2018.

[Jin et al., 2020a] Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. Self-supervised learning on graphs: Deep insights and new direction. arXiv preprint arXiv:2006.10141, 2020.

[Jin et al., 2020b] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In ICML, 2020.

[Jin et al., 2020c] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In ICML, 2020.

[Kim et al., 2019] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Sijian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thissen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. Nucleic acids research, 2019.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In ICLR, 2017.

[Klicpera et al., 2019] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directational message passing for molecular graphs. In ICLR, 2019.

[Klicpera et al., 2021a] Johannes Klicpera, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. NeurIPS, 2021.

[Klicpera et al., 2021b] Johannes Klicpera, Chandan Yeshwanth, and Stephan Günnemann. Directional message passing on molecular graphs via synthetic coordinates. NeurIPS, 2021.

[Krenn et al., 2020] Mario Krenn, Florian Häse, Akshat Kumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. Machine Learning: Science and Technology, 2020.

[Kusner et al., 2017] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In ICML, 2017.

[Landrum, 2020] G. A. Landrum. Rdkit: Open-source cheminformatics software. http://www.rdkit.org, 2020. Accessed: 2023-01-01.

[Li et al., 2016] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated graph sequence neural networks. In ICLR, 2016.

[Li et al., 2022] Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In AAAI, 2022.

[Lin et al., 2020] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In IJCAI, 2020.

[Lin et al., 2020a] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In IJCAI, 2020.
[Liu et al., 2022a] Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph rationalization with environment-based augmentations. In *KDD*, 2022.

[Liu et al., 2022b] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *ICLR*, 2022.

[Liu et al., 2022c] Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuwiang Ji. Spherical message passing for 3d molecular graphs. In *ICLR*, 2022.

[Lowe, 2012] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[Lyu et al., 2021] Tengfei Lyu, Jianliang Gao, Ling Tian, Zhao Li, Peng Zhang, and Ji Zhang. Mdmn: A multimodal deep neural network for predicting drug-drug interaction events. In *IJCAI*, 2021.

[Ma et al., 2018] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. Drug similarity integration through attentive multi-view graph auto-encoders. In *IJCAI*, 2018.

[Maziarz et al., 2022] Krzysztof Maziarz, Henry Richard Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefel, Marwin Segler, and Marc Brockschmidt. Learning to extend molecular scaffolds with structural motifs. In *ICLR*, 2022.

[Rong et al., 2020] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *NeurIPS*, 2020.

[Saebi et al., 2023] Mandana Saebi, Bozhao Nan, John E Herr, Jessica Wahlers, Zhichun Guo, Andrezej M Zurawiski, Thierry Kogej, Per-Ola Norby, Abigail G Doyle, Nitesh V Chawla, et al. On the use of real-world datasets for reaction yield prediction. *Chemical Science*, 2023.

[Shi et al., 2020] Chence Shi, Minkai Xu, Zhaoceng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *ICLR*, 2020.

[Shi and Karypis, 2020] Zeren Shui and George Karypis. Heterogeneous molecular graph neural networks for predicting molecular properties. In *ICDM*, 2020.

[Stark et al., 2022] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gns for molecular property prediction. In *ICML*, 2022.

[Sterling and Irwin, 2015] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 2015.

[Sun et al., 2020] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2020.

[Sun et al., 2021] Mengying Sun, Jing Xing, Huijun Wang, Bin Chen, and Jiayu Zhou. Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge. *arXiv preprint arXiv:2106.04509*, 2021.

[Tang et al., 2020] Bowen Tang, Skyler T Kramer, Meijuan Fang, Yingkun Qiu, Zhen Wu, and Dong Xu. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of cheminformatics*, 2020.

[Veličković et al., 2018] Petar Veličković, Guillem Cucurull, Aran-txa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[Wang et al., 2021] Yaqing Wang, Abulikemu Abuduweili, Quan-ning Yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. *NeurIPS*, 2021.

[Wang et al., 2022a] Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kung-hyun Cho, Heng Ji, Jiawei Han, and Martin Burke. Chemical-reaction-aware molecule representation learning. In *ICLR*, 2022.

[Wang et al., 2022b] Limei Wang, Yi Liu, Yuchao Lin, Haoran Liu, and Shuwiang Ji. Comenet: Towards complete and efficient message passing for 3d molecular graphs. In *NeurIPS*, 2022.

[Weininger et al., 1989] David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 1989.

[Wu et al., 2018] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Anees S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 2018.

[Xu et al., 2019] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

[Xu et al., 2021] Minkai Xu, Wujie Wang, Shitong Luo, Chence Shi, Yoshua Bengio, Rafael Gomez-Bombarelli, and Jian Tang. An end-to-end framework for molecular conformation generation via bilevel programming. In *ICML*, 2021.

[Yang et al., 2021a] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *IJCAI*, 2021.

[Yang et al., 2021b] Shuwen Yang, Ziyao Li, Guojie Song, and Lingsheng Cai. Deep molecular representation learning via fusing physical and chemical information. *NeurIPS*, 2021.

[Yang et al., 2022] Nianzuo Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. In *NeurIPS*, 2022.

[Yang et al., 2023] Nianzuo Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *WWW*, 2023.

[You et al., 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *NeurIPS*, 2020.

[Zhang et al., 2019] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *KDD*, 2019.

[Zhang et al., 2021] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *NeurIPS*, 2021.

[Zhu et al., 2022] Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Unified 2d and 3d pre-training of molecular representations. In *KDD*, 2022.