Journal Title Here, 2022, pp. 1–15

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

A Systematic Survey of Molecular Pre-trained Models

Jun Xia,[1,2] Yanqiao Zhu,[3] Yuanqi Du,[4] Yue Liu[2] and Stan Z. Li[2,*]

[1]Zhejiang University, No. 388 Yuhangtang Road, Zhejiang Province, P. R. China, 310058, [2]AI Division, Westlake University, No. 600 Dunyu Road, Zhejiang Province, P. R. China, 310030, [3]University of California at Los Angeles, 404 Westwood Plaza, Los Angeles, CA, United States, 90095 and [4]Cornell University, 402 Gates Hall, Ithaca, NY, United States, 14850
[*]Corresponding author: Stan.ZQ.Li@westlake.edu.cn; +86 0571-85128506
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Obtaining effective molecular representations is at the core of a series of important chemical tasks ranging from property prediction to drug design. So far, deep learning has achieved remarkable success in learning representations for molecules through automated feature learning in a data-driven fashion. However, training deep neural networks from scratch often requires sufficient labeled molecules which are expensive to acquire in real-world scenarios. To alleviate this issue, inspired by the success of the pretrain-then-finetune paradigm in natural language processing, tremendous efforts have been devoted to Molecular Pre-trained Models (MPMs), where neural networks are pre-trained using large-scale unlabeled molecular databases and then fine-tuned for diverse downstream tasks. Despite the prosperity, this field is fast-growing and a systematic roadmap is urgently needed for both methodology advancements and practical applications in both machine learning and scientific communities. To this end, this paper provides a systematic survey of pre-trained models for molecular representations. Firstly, to motivate MPMs studies, we highlight the limitations of training deep neural networks for molecular representations. Next, we systematically review recent advances on this topic from several key perspectives including molecular descriptors, encoder architectures, pre-training strategies, and applications. Finally, we identify several challenges and discuss promising future research directions.

Key words: Molecular Representation Learning, Pre-trained Models, Molecular Property Prediction, Self-supervised learning, Drug-Drug Interaction, Drug-Target Interaction

1. Introduction

Extracting vector representations for molecules is critical to applying machine learning methods for a broad spectrum of molecular tasks [101]. Initially, molecular fingerprints are developed to encode molecules into binary vectors with rule-based algorithms [13, 68, 79]. Later, with the success of deep neural networks, a variety of methods have been employed to encode molecular descriptors in a task-driven fashion [3, 12, 20, 43, 45, 71]. Early attempts exploit sequence-based neural architectures (RNNs, LSTMs, and Transformers) to encode molecules represented in Simplified Molecular-Input Line-Entry System (SMILES) strings [10, 105]. Later, it is argued that molecules can be naturally represented in graph structures with atoms as nodes and bonds as edges. This inspires a line of works to leverage this structured inductive bias for better molecular representations [46, 83, 86, 87, 116, 117]. Specifically, the key advancements underneath their approaches are Graph Neural Networks (GNNs) or Message-Passing Neural Networks (MPNNs), which take into account graph structures and other attributive features simultaneously by recursively aggregating node features from neighborhoods [51, 103]. It is noted that this recursive way of learning graph representations ensembles with the traditional fingerprint representations to some extent [20]. Recently, there is another line of development of GNNs for molecular representations that models 3D geometric symmetries of molecular conformation, considering the molecules are in a constant motion in 3D space by nature [17, 61, 85, 88].

Most of the aforementioned works specifically focus on learning molecular representations under supervised settings. This learning paradigm often requires sufficient labeled molecular data, which impedes their wider usage in practice due to the following two reasons:

- *Scarcity of labeled data:* Task-specific labels of molecules can be extremely scarce because data labeling often requires time-consuming and resource-costly wet-lab experiments.

- *Poor out-of-distribution generalization:* Learning molecules with different sizes or functional groups requires out-of-distribution generalization in many real-world scenarios. For example, when one wants to predict the chemical properties of a newly synthesized molecule, which is different from all the molecules in the training set. However, it is observed that current neural networks cannot well extrapolate to out-of-distribution molecules [38, 40, 44].

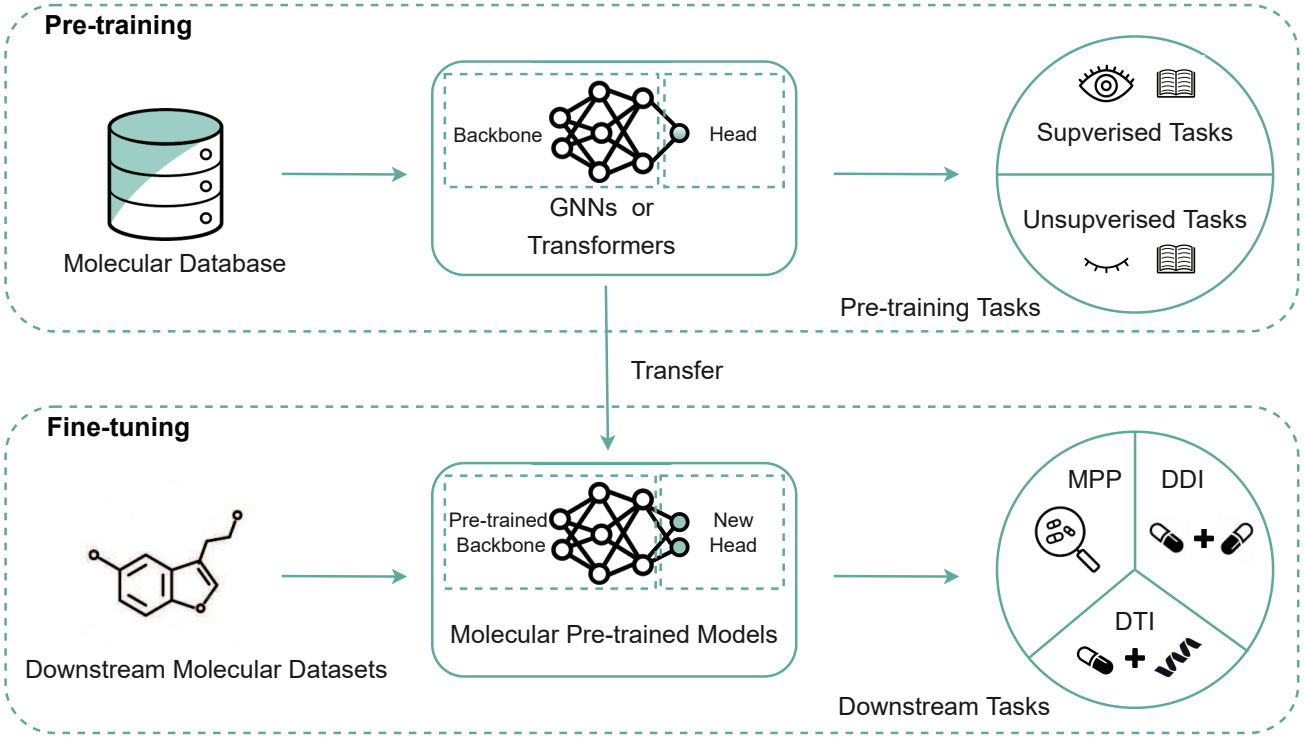


Fig. 1. The pre-training–fine-tuning paradigm for Molecular Pre-trained Models (MPMs): Firstly, a neural encoder is pre-trained with unsupervised objectives to obtain general knowledge of molecular data. Then, the pre-trained model is fine-tuned for various downstream tasks. Backbone: feature extraction networks that compute representations from the input. Head: a simple neural network that transforms the representations to the output space. MPP: Molecular Property Prediction; DDI: Drug-Drug Interaction; DTI: Drug-Target Interaction.

Inspired by the recent success of Pre-trained Language Models (PLMs) in Natural Language Processing (NLP), MPMs have been introduced to learn universal molecular representations from massive unlabeled molecules and fine-tuned over downstream tasks with task-specific labeled data. An illustrative pipeline is shown in Figure 1. In the beginning, researchers adopt sequence-based pre-training strategies on string-based molecular data such as SMILES. A typical strategy is to pre-train the neural encoders to predict randomly masked tokens like BERT [16]. This line of works include ChemBERTa [10], SMILES-BERT [105], Molformer [81], etc. More recently, the community explores pre-training on (both 2D and 3D) molecular graphs. Specifically, pre-training on 2D graphs focuses on exploiting the structural information of the graph topology, while pre-training on 3D graphs takes advantage of 3D conformational information. For example, Hu et al. [38] propose two pre-training strategies, (1) masking atom or edge attributes and predicting the masked attributes and (2) predicting the context subgraph obtained by taking the K-hop neighbors around a center atom. You et al. [125] propose to maximize the agreements between paired molecular graph augmentations using a contrastive objective [9]. Zaidi et al. [128] demonstrate that performing denoising on the conformational space can be helpful for learning molecular force fields.

Although MPMs have been increasingly applied in molecular representation learning, a review of this fast-growing field is still lacking. To help audiences of diverse backgrounds to understand, use, and develop MPMs for various practical tasks, we present a systematic survey that reviews the current progress of MPMs for molecules. The contributions of this work can be summarized from the following four aspects:

- *Structured taxonomy.* As shown in Figure 2, we contribute a structured taxonomy to provide a broad overview of the field, which categorizes existing works from four perspectives: molecular descriptors, neural architectures, pre-training strategies, and applications.
- *Current progress.* According to the taxonomy, we systematically delineate the current research directions on the topic of pre-trained models for molecules.
- *Abundant resources.* We collect abundant resources including open-sourced MPMs, available datasets, and an important paper list at https://github.com/junxia97/awesome-pretrain-on-
We promise to continuously update these resources.
- *Future directions.* We discuss the limitations of existing works and suggest several promising future research directions.

2. Molecular Descriptors

In order to input molecules to machine learning models, molecules have to be featurized in numerical descriptors. A variety of molecular descriptors including binary vectors of structure fragments, sequences of characters, molecular graphs, and molecular geometries are designed to describe molecules in a concise format. However, different descriptors may result in shifted focuses on certain molecular information. In this section, we briefly review these molecular descriptors.

Table 1. A summary of molecular features in the basic feature setting.

Attribute	Dimension	Explanation
Atom type	118	Type of atom (e.g., C, N, O), by atomic number
Chirality	4	Unspecified, tetrahedral cw, tetrahedral ccw, or other
Bond type	4	Single, double, triple, or aromatic
Bond direction	3	Endupright, enddownright, or other

Table 2. A summary of molecular features in the rich feature setting.

Attribute	Dimension	Explanation
Atom type	100	Type of atom (e.g., C, N, O), by atomic number
Formal charge	5	Integer electronic charge assigned to atom
Number of bonds	6	Number of bonds the atom is involved in
Chirality	5	Unk, unspecified, tetrahedral-CW, tetrahedralL-CCW, or other
Number of H	5	Number of bonded hydrogen atoms
Atomic mass	1	Mass of the atom, divided by 100
Aromaticity	1	Whether this atom is part of an aromatic system
Hybridization	5	sp, sp2, sp3, sp3d, or sp3d2
Bond type	4	Single, double, triple, or aromatic
Stereo	6	None, any, E/Z or cis/trans
In ring	1	Whether the bond is part of a ring
Conjugated	1	Whether the bond is conjugated

Fingerprints. Molecular fingerprints describe the presence or absence of particular substructures of a molecule with binary strings. For example, PubChemFP [108] encodes 881 structural key types that correspond to the substructures for a fragment of all compounds in the PubChem database. Morgan fingerprints [68] assign numeric identifiers to each atom and iteratively update these atom descriptors among neighboring atoms using a hash function.

Sequences. The most frequently-used sequential descriptor for molecules is the Simplified Molecular-Input Line-Entry System (SMILES) [111] owing to its versatility and interpretability. Each atom is represented as a respective ASCII symbol. Chemical bonds, branching, and stereochemistry are denoted by specific symbols in SMILES strings. However, large fractions of SMILES strings do not correspond to valid molecules. As a remedy, SELF-referencing Embedded Strings (SELFIES) [54], a string-based descriptor of molecules, is developed recently to tackle this issue, such that every SELFIES string denotes a valid molecule.

2D graphs. Molecules can be represented as 2D graphs naturally, with atoms as nodes and bonds as edges. Moreover, each node and edge can also carry informative feature vectors denoting the atom/bond types for instance. Even though 2D graphs turn out to be a natural descriptor for molecules, this kind of descriptor has several limitations. For example, for two molecular graphs that differ in chirality, most mainstream graph neural networks cannot distinguish them when the atom and bond types serve as the sole features. To enrich the graph representations with more meaningful features, two feature sets are commonly included: *basic feature* and *rich feature*. The former is a minimal set of node and bond features that unambiguously describe the two-dimensional structure of molecules [38] and the latter includes additional atom features such as aromaticity and hybridization, and bond features such as ring information [80]. We summarize these two feature sets Table 1 and Table 2, respectively.

3D graphs. 3D molecular geometries represent the spatial arrangements of each atom in the 3D space. Specifically, it includes a list of atoms with atom types and atomic coordinates. Different from 2D molecular graphs which focus on topological information, 3D geometries encode conformational information, which is critical to many molecular properties, especially quantum properties. In addition, it is also possible to directly infer chirality given 3D molecular geometries in the 3D space.

3. Neural Encoder Architectures

In this section, we review neural architectures for encoding molecular descriptors into continuous vectors. We specifically discuss two main architectures, transformers and graph neural networks due to their wide popularity in the open literature.

3.1. Transformers

Transformer [102] has become the de facto standard for pre-training language models with unlabeled datasets, owing to its powerful self-attention mechanism. Also, transformers can be used for effectively embedding all four aforementioned molecular descriptors (fingerprints, sequences, 2D graphs, and 3D graphs).

Transformers for fingerprints/sequences. Sequential descriptors share a similar intrinsic structure with text data. Since Transformers are powerful for processing sequential data and modeling the complex relationships among each token of the input sequence, we can split sequence-based molecular descriptions into a series of tokens indicating the atoms and bonds at first and then directly apply Transformers on top of these tokens. In the existing literature, Transformers have been adopted as neural encoders to pre-train molecular representations with SMILES strings [10, 34, 81, 105]. More recently, a self-attention based neural encoder is devised to encode the molecular fingerprints for molecular pre-training as well [135].

Transformers for 2D/3D molecular graphs. To exploit the topology and geometry information of molecules, tremendous efforts have been devoted to graph Transformers which receive molecular graphs as input as well. For example, GROVER [80]

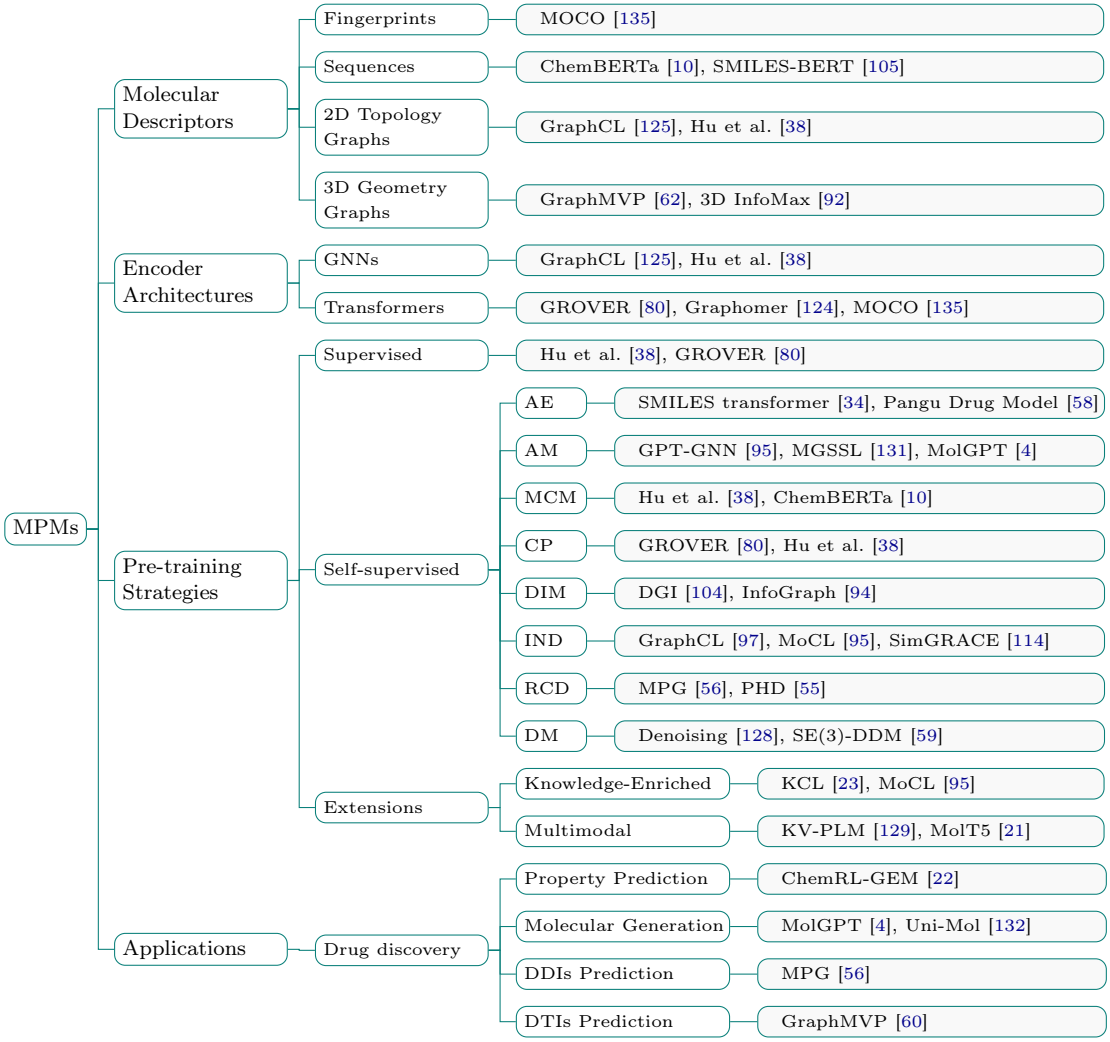


Fig. 2. The taxonomy of Molecular Pre-trained Models (MPMs) with representative examples.

and MPG [56] first utilize Graph Neural Networks (GNNs) to capture local structural information of molecular graphs and then feed the learned embeddings into a sequence of Transformer layers. Later works attempt to incorporate graph structural information directly into the Transformer architecture with improved positional encodings [7, 41, 67] and/or improved attention maps from graph topology [124]. Additionally, Zhou et al. [132] improve graph Transformers by injecting geometry information as the inter-atomic distance into attention maps.

3.2. Graph Neural Networks (GNNs)

Apart from Transformers, Graph Neural Networks (GNNs) or Message-Passing Neural Networks (MPNNs) can also be used for encoding structural information for graph-structured data. Specifically, these models explicitly consider the interactions of connected nodes by passing messages between them, which naturally fulfills permutation equivariance such that the output node embeddings follow the input node ordering. Since molecules can be represented by graph structures, GNNs are a natural choice for learning 2D molecular graph representations. Recent works also explore extending GNNs to 3D geometries as well. In this case, complex convolutional layers are developed to allow the node representations to follow certain physical rules, such as equivariance to E(3) transformations.

2D graph neural networks. 2D GNNs incorporate topology structures by iteratively performing aggregation of local neighboring information, which creates a contextual representation for each node. Further, graph-level representations can be obtained by taking a pooling operation over all the nodes [66]. The common GNN architectures for 2D graphs include GCN [49], GraphSAGE [29], GAT [103], GIN [119], etc.

3D graph neural networks. In order to consider the additional conformation structures of molecular graphs, a class of 3D GNNs have been proposed to incorporate geometry information in the message-passing framework. Specifically, Schütt et al. [86] propose a continuous-filter convolution layer to encode molecular geometry information represented by interatomic distances. Subsequent works [17, 85, 100] design 3D GNNs which are equivariant to translations and rotations and improve the data efficiency by removing the unneeded degree of freedom. Some other works [53, 61] introduce additional angular information to improve the expressiveness of the 3D GNNs.

Table 3. Notions and loss functions of supervised and self-supervised pre-training strategies. $\mathcal{M}$, $\mathcal{V}$ denote the molecules and the atom set, respectively. $s(\cdot,\cdot)$ denotes a similarity metric. $\mathcal{D}$ is the pre-training molecular dataset.

Task	Loss Function	Description
Supervised	$\mathcal{L}_{\text{Supervised}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p(\mathcal{Y} \mid \mathcal{M})$	Supervised pre-training. $\mathcal{Y}$ is the given label.
AE	$\mathcal{L}_{\text{AE}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p(\widehat{\mathcal{M}} \mid \mathcal{M})$	Molecule reconstruction. $\widehat{\mathcal{M}}$ denotes the reconstructed one.
AM	$\mathcal{L}_{\text{AM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \Sigma_{i=1}^{\lvert\mathcal{V}\rvert} \log p(\mathcal{X}_i, \mathcal{E}_i \mid \mathcal{X}_{<i}, \mathcal{E}_{<i})$	$\mathcal{X}_{<i}, \mathcal{E}_{<i}$ are the atoms and bonds attributes generated before node i in molecule $\mathcal{M}$.
MCM	$\mathcal{L}_{\text{MCM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \Sigma_{\widetilde{\mathcal{M}}\in m(\mathcal{M})} \log p(\widetilde{\mathcal{M}} \mid \mathcal{M}_{\setminus m(\mathcal{M})})$	$m(\mathcal{M})$ are the masked components from $\mathcal{M}$ and $\mathcal{M}_{\setminus m(\mathcal{M})}$ are the rest.
CP	$\mathcal{L}_{\text{CP}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p(t \mid \mathcal{M}_1, \mathcal{M}_2)$	$t = 1$ if neighborhood components $\mathcal{M}_1$ and contexts $\mathcal{M}_2$ share the same center atom in $\mathcal{M}$.
IND	$\mathcal{L}_{\text{IND}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}[\log s(\mathcal{M}, \mathcal{M}') - \log \Sigma_{\mathcal{M}^-\in\mathcal{N}} s(\mathcal{M}, \mathcal{M}^-)]$	$\mathcal{N}$ is a set of negatives (other molecules); $\mathcal{M}'$ is a positive sample (e.g., the augmented version).
DIM	$\mathcal{L}_{\text{DIM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}[\log s(\mathcal{M}, \mathcal{C}) - \log \Sigma_{\mathcal{C}^-\in\mathcal{N}} s(\mathcal{M}, \mathcal{C}^-)]$	$\mathcal{N}$ is a set of negatives; $\mathcal{C}$ is a substructure of $\mathcal{M}$; $\mathcal{C}^-$ is a substructure of the other molecule.
RCD	$\mathcal{L}_{\text{RCD}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}[\log p(t \mid \mathcal{M}_1, \mathcal{M}_2)]$	$t = 1$ if two half molecules $\mathcal{M}_1$ and $\mathcal{M}_2$ are homologous couples from $\mathcal{M}$.
DM	$\mathcal{L}_{\text{DM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p(\mathcal{M} \mid \mathcal{M} + \mathcal{R})$	$\mathcal{R}$ is the random noise imposed on the input molecule.

4. Pre-training Strategies

Existing pre-training strategies broadly fall into two categories: supervised pre-training and self-supervised pre-training. Supervised pre-training leverages easily accessible annotations as supervision signals. For example, Hu et al. [38] propose to pre-train GNNs to predict the properties of all the available molecules that have been experimentally measured. However, supervised pre-training tasks do not always align well with downstream tasks, thus negative transfer could occur. In contrast, the self-supervised pre-training is a more plausible alternative for molecules, which we will elaborate on in the following subsections. We summarize and formulate several representative pre-training strategies using a unified symbolic system in Table 3.

4.1. AutoEncoders (AE)

Reconstructing molecules with autoencoders (Fig. 3a) serves as a natural self-supervised target for learning discriminative molecular representations. The prediction targets in molecule reconstructions are (partial) structures of the given molecules such as the attribute of a subset of atoms or chemical bonds. For example, Honda et al. [34] build a transformer-based encoder-decoder network to reconstruct the molecules represented by SMILES strings. Inspired by the success of autoencoders in other domains [48], graph autoencoders (GAEs) [50] have also been widely adopted to pre-train on molecular graphs via reconstructing the adjacency matrix of the original molecular graphs. Formally, the pre-training objective can be defined as

$$\mathcal{L}_{\text{AE}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p(\widehat{\mathcal{M}} \mid \mathcal{M}), \tag{1}$$

where $\mathcal{M}$ and $\widehat{\mathcal{M}}$ denote the original molecule and the reconstructed molecule, respectively. $\mathcal{D}$ is the pre-training molecular dataset. More recently, unlike conventional autoencoders with the same data types for the input and output, Lin et al. [58] pre-train a graph-to-sequence asymmetric conditional variational autoencoders to obtain molecular representations. Although autoencoders can learn meaningful representations for molecules, they fail to capture the inter-molecule relationships, which limits their performance.

4.2. Autoregressive Modeling (AM)

Autoregressive modeling factorizes the molecular input as a sequence of sub-components and then it predicts the sub-components one by one conditioning on previous sub-components in the sequence. Following the idea of GPT [6] which achieves great success in pre-training language models, MolGPT [4] pre-trains a transformer network to predict the next token in the SMILES strings autoregressively. For molecular graphs, GPT-GNN [39] reconstructs the molecular graph in a sequence of steps (Fig. 3b), in contrast to graph autoencoders that reconstruct the whole graph 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 maximizes the likelihood of the node and edges generated in each iteration. Then, it iteratively generates nodes and edges until all masked nodes are generated. Formally, its autoregressive modeling objective is

$$\mathcal{L}_{\text{AM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \Sigma_{i=1}^{\lvert\mathcal{V}\rvert} \log p(\mathcal{X}_i, \mathcal{E}_i \mid \mathcal{X}_{<i}, \mathcal{E}_{<i}), \tag{2}$$

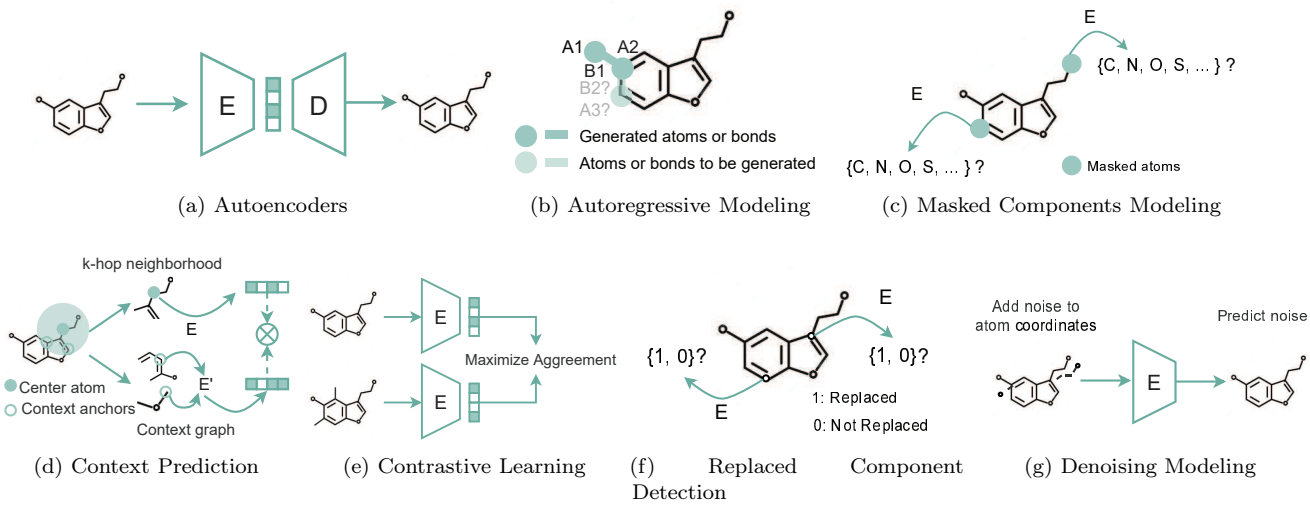
where $\mathcal{X}_{<i}, \mathcal{E}_{<i}$ are the atoms and bonds attributes generated before the index i in the molecule $\mathcal{M}$. Analogously, MGSSL [131] generates molecular graph motifs instead of individual atoms or bonds. Compared with other pre-training strategies, MPMs by autoregressive modeling are capable of completing molecules, which is conducive to molecule generation. However, the autoregressive pre-training procedure is computationally more expensive.

4.3. Masked Components Modeling (MCM)

Similar to Masked Language Modeling (MLM) which randomly masks out tokens from the input sentences and then trains the model to predict the masked tokens with the rest of the tokens [16], MCM (Fig. 3c) first masks out some components (e.g., atoms, bonds, subgraphs, etc.) of the molecules and then trains the model to predict them. The pre-training objective of MCM can be formulated as

$$\mathcal{L}_{\text{MCM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \Sigma_{\widetilde{\mathcal{M}}\in m(\mathcal{M})} \log p(\widetilde{\mathcal{M}} \mid \mathcal{M}_{\setminus m(\mathcal{M})}), \tag{3}$$

where $m(\mathcal{M})$ are the masked components from the molecule $\mathcal{M}$ and $\mathcal{M}_{\setminus m(\mathcal{M})}$ are the remaining components. For sequence-based pre-training, following the masked language modeling in NLP, ChemBERTa [10], SMILES-BERT [105], and Molformer [81] mask some characters in the SMILES or SELFIES strings, and then recover the masked characters based on the output of the transformer for the input corrupted SMILES or SELFIES strings. For molecular graph pre-training, Hu et al. [38] propose attribute masking where the input atom/chemical bond attributes are randomly masked, and the GNN is pre-trained to predict them. Also, GROVER [80] attempts to predict the masked subgraphs to capture the contextual information in the molecular graphs. These masking methods are especially beneficial for richly-annotated molecular graphs. For example, masking node attributes (atom types) enables GNNs to learn simple chemistry rules such as valency,



(a) Autoencoders (b) Autoregressive Modeling (c) Masked Components Modeling

(d) Context Prediction (e) Contrastive Learning (f) Replaced Component Detection (g) Denoising Modeling

Fig. 3. Semantic diagrams of various unsupervised pre-training strategies. E: encoder; D: decoder.

as well as potentially more complex chemistry phenomena such as the electronic or steric effects of functional groups. Additionally, compared with the autoregressive modeling we describe in Section 4.2, MCM predicts the masked components (atoms/bonds) based on their surrounding environments while autoregressive modeling predicts them only dependent on the components appearing before them in the sequence. As a result, MCM allows the MPMs to comprehensively capture chemical rules. However, the input of pre-training models with MCM often contains artificial symbols that never occur in downstream tasks, which creates a gap between pre-training and fine-tuning stages [38, 122]. This critical issue remains unsolved for molecular pre-training.

4.4. Context Prediction (CP)

Context prediction (Fig. 3d) is proposed to capture the semantics of molecules/atoms in a context-aware manner. For example, Hu et al. [38] use subgraphs in molecules to predict their surrounding context structures with GNNs. GROVER [80] attempts to predict the context-aware properties of the target atoms/bonds within some local subgraphs. Here, the properties are some atom-bond count terms around the target atom/bond. Although effective, CP [38] requires an auxiliary neural model to encode the context into a fixed vector, which introduces more computational overhead for large-scale pre-training. Formally, the general formulation of CP is

$$\mathcal{L}_{\mathrm{CP}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p\left(t \mid \mathcal{M}_1, \mathcal{M}_2\right), \tag{4}$$

where $t = 1$ if neighborhood components $\mathcal{M}_1$ and contexts $\mathcal{M}_2$ share the same center atom and otherwise $t = 0$.

4.5. Contrastive Learning (CL)

Contrastive pre-training has emerged as one of the most popular strategies for molecular representations. According to the contrastive granularity (e.g., molecule level or subgraph level), CL can be categorized into two categories: Deep InfoMax (DIM) and Instance Discrimination (IND).

Deep InfoMax (DIM) Deep InfoMax is originally proposed for images to learn the representation by maximizing the mutual information between an image representation and local regions of the image [33]. For molecular graphs, InfoGraph [94] firstly proposes to obtain expressive molecular representations via maximizing the mutual information between molecule- and substructure-level representations of different granularity, which can be formally described as

$$\mathcal{L}_{\mathrm{DIM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}[\log s(\mathcal{M}, \mathcal{C}) - \log \Sigma_{\mathcal{C}^-\in\mathcal{N}} s(\mathcal{M}, \mathcal{C}^-)], \tag{5}$$

where $\mathcal{N}$ is a set of negative samples, $\mathcal{C}$ is a substructure of $\mathcal{M}$, $\mathcal{C}^-$ is a substructure of the other molecule, and $s(\cdot, \cdot)$ denotes a similarity metric. Follow-up work MVGRL [31] performs node diffusion to generate an augmented molecular graph and then maximizes the mutual information between original and augmented views by contrasting atom representations of one view with molecular representations of the other view and vice versa. For molecular geometry, 3DInfoMax [92] maximizes the mutual information between the learned 3D geometry and 2D graph representations.

Instance Discrimination (IND) Instance Discrimination (Fig. 3e) is one of the most popular pre-training strategies which aims to learn molecular representations by pushing the augmented molecule close to the anchor molecule (positive pairs) but away to other molecules (negative pairs). For molecular representations, GraphCL [125] and its variants [23, 72, 95, 97, 109, 110, 118, 126] propose various advanced augmentation strategies for molecule-level pre-training represented by graphs. More recently, some works attempt to simplify the above contrastive pre-training framework by discarding the negative pairs [99], parameterized mutual information estimator [130], or even molecular graph data augmentations [114], respectively. Additionally, some recent works maximize the agreements between the different descriptors of identical molecules and repel the different ones. Specifically, SMICLR [72] jointly pre-trains a graph encoder and a SMILES string encoder to perform the contrastive learning objective; MM-Deacon [28] utilizes two separate Transformers to encode the SMILES and the International Union of Pure and Applied Chemistry (IUPAC) of molecules to expressive representations, after which the contrastive objectives are used to maximize mutual information for

SMILES and IUPAC from the same molecule and distinguish SMILES and IUPAC from different molecules; GeomGCL [57] adopts a dual-view Geometric Message Passing Neural Network (GeomMPNN) to encode both 2D and 3D graphs of a molecule and design geometric graph contrastive objective. The general formulation of the IND pre-training objective is

$$\mathcal{L}_{\text{IND}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}\left[\log s(\mathcal{M},\mathcal{M}') - \log \Sigma_{\mathcal{M}^-\in\mathcal{N}} s(\mathcal{M},\mathcal{M}^-)\right]. \tag{6}$$

Although molecular contrastive pre-training has achieved promising results, several critical issues impede its broader applications. (1) It is difficult to preserve semantics during molecular augmentations. Existing solutions pick suitable augmentations with manual trial-and-errors [125], cumbersome optimization [126], or through the guidance of expensive domain knowledge [95], which are sub-optimal. (2) Pulling similar molecules closer may not always hold true for molecular contrastive learning. For example, in cases of molecular activity cliffs [93], similar molecules hold completely different properties, and the contrastive objective that still pulls their embeddings closer might not be appropriate. It remains to be explored whether there are more suitable augmentations or augmentations-free frameworks for molecules pre-training. (3) The contrastive framework pushes away all the other molecular graphs regardless of their true semantics, which will undesirably repel the molecules of similar properties and undermine the performance due to the false negative issue advocated in a recent work [115].

4.6. Replaced Component Detection (RCD)

Replaced component detection (Fig. 3f) is proposed as an effective pre-training task with random permutations of input molecules. Specifically, PHD [55] decomposes molecules into two parts and permutes the molecular structures by random combinations of individual parts from different molecules. Then the neural encoder is trained to classify whether the two components are from the same molecule. The objective function is

$$\mathcal{L}_{\text{RCD}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}\left[\log p(t \mid \mathcal{M}_1, \mathcal{M}_2)\right], \tag{7}$$

where $t = 1$ if two half molecules $\mathcal{M}_1$ and $\mathcal{M}_2$ are from the same molecule $\mathcal{M}$ and otherwise $t = 0$.

Although RCD can help MPMs capture intrinsic patterns underlying the molecular structures, it is essentially a binary classification task, which is less challenging than MCM that we introduce in Section 4.3. Consequently, the pre-training process will converge to a high value quickly with this simple pre-training task, which consequently makes the pre-trained models capture less transferable knowledge and impairs the generalization or adaptation to novel tasks [11, 78].

4.7. Denoising Modeling (DM)

Inspired by Noisy Nodes [26] that incorporates denoising (Fig. 3g) as an auxiliary task to improve performance, a recent work [128] adds noise to atomic coordinates of 3D geometry and pre-trains the encoders to predict the noise. They further demonstrate that denoising objective in molecular pre-training is approximately equivalent to learning a molecular force field, shedding light on how denoising aids molecular pre-training. Concurrently, considering that the masked atom types can be inferred with 3D atomic positions, the vanilla masked atom type prediction task can be extremely simple [132]. As a remedy, Uni-Mol [132] designs a 3D position denoising pre-training task, where noises are added to the atomic coordinates to make the pre-training task of masked atom prediction (see section 4.3) more challenging, and therefore encourage the model to learn more transferable knowledge. The general pre-training objective of DM is

$$\mathcal{L}_{\text{DM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p(\mathcal{M} \mid \mathcal{M} + \mathcal{R}), \tag{8}$$

where $\mathcal{R}$ denotes the random noise imposed on the input molecule $\mathcal{M}$.

4.8. Extensions

4.8.1. Knowledge-Enriched Pre-training

MPMs usually learn general molecular representations from the large molecular database. However, they often lack domain-specific knowledge. To enhance their performance, several recent works try to inject external knowledge into MPMs. For example, GraphCL [125] first points out that bond perturbations (adding or dropping the bonds as data augmentations) are conceptually incompatible with domain knowledge and empirically unhelpful for contrastive pre-training on chemical compounds. Therefore, they avoid adopting bond perturbations for molecular graph augmentation. To incorporate the domain knowledge into pre-training more explicitly, MoCL [95] proposes a new molecular augmentation operator called substructure substitution, in which a valid substructure of a molecule is replaced by a bioisostere [65] which produces a new molecule with similar physical or chemical properties as the original one. More recently, to capture the correlations between atoms that have common attributes but are not directly connected by chemical bonds, KCL [23] constructs a chemical element Knowledge Graph (KG) to summarize microscopic associations between chemical elements and contributes a novel Knowledge-enhanced Contrastive Learning (KCL) framework for molecular representation learning. Additionally, MGSSL [131] first leverages the BRICS algorithm [15] to derive semantically meaningful motifs and then pre-trains neural encoders to predict the motifs in an autoregressive manner. ChemRL-GEM [22] proposes to utilize molecular geometry information to enhance molecular graph pre-training. It designs a geometry-based graph neural network 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. Zhu et al. [135] propose to maximize the consistency between the view embeddings of four molecular descriptors and their aggregated embedding using a contrastive objective. Therefore, the various descriptors can complement each other for the molecular property prediction tasks. Although knowledge-enhanced pre-training helps GMs capture chemical domain knowledge, it requires expensive prior guidance, which poses a hurdle to broader applications when the prior is incomplete or incorrect.

4.8.2. Multimodal Pre-training

In addition to the descriptors mentioned in Section 2, molecules can also be described using other modalities including images and biochemical texts. Inspired by the advances of multimodal pre-training in computer vision and NLP domains [74, 77], some recent works perform multimodal pre-training on molecules. For example, KV-PLM [129] first tokenizes both the SMILES strings and biochemical texts. Then, they randomly mask part of the tokens and pre-train the neural encoders to

Table 4. List of representative MPMs.

	Model	Input	Architecture	Pre-training Task	Pre-training Database	#Params.	Access Link
SEQUENCE	SMILES Transformer [34]	SMILES	Transformer	AE	ChEMBL (861k)	—	Link
	ChemBERTa [10]	SMILES/SELFIES	Transformer	MCM	PubChem (77M)	—	Link
	SMILES-BERT [105]	SMILES	Transformer	MCM	ZINC15(∼ 18.6M)	—	Link
	Molformer [81]	SMILES	Transformer	MCM	ZINC15(1B) + PubChem(111M)	—	—
GRAPH & GEOMETRY	Hu et al. [38]	Graph	5-layer GIN	CP + MCM	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	GraphCL [125]	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	JOAO [126]	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	AD-GCL [97]	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	GraphLog [120]	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	MGSSL [131]	Graph	5-layer GIN	MCM + GAM	ZINC15 (250K)	∼ 2M	Link
	MPG [56]	Graph	MolGNet [56]	RCD + MCM	ZINC + ChEMBL (11M)	53M	Link
	LP-Info [127]	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	SimGRACE [114]	Graph	5-layer GIN	IND	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	GraphMAE [37]	Graph	5-layer GIN	AE	ZINC15(2M) + ChEMBL(456K)	∼ 2M	Link
	GROVER [80]	Graph	GTransformer [80]	CP + MCM	ZINC + ChEMBL (10M)	48M∼100M	Link
	MolCLR [106]	Graph	GCN + GIN	IND	PubChem (10M)	—	Link
	Graphomer [124]	Graph	Graphomer [124]	Supervised	PCQM4M-LSC (∼3.8M)	—	Link
	Denoising [128]	Geometry	GNS [84]	DM	PCQM4Mv2(∼3.4 M)	—	Link
MULTIMODAL/KNOWLEDGE-ENRICHED	DMP [133]	Graph + SMILES	DeeperGCN + Transformer	MCM + IND	PubChem (110M)	104.1 M	Link
	GraphMVP [60]	Graph + Geometry	5-layer GIN + SchNet [86]	IND + AE	GEOM (50k)	∼ 2M	Link
	3D Infomax [92]	Graph + Geometry	PNA [14]	IND	QM9(50K) + GEOM(140K) + QMugs(620K)	—	Link
	KCL [23]	Graph + KG	GCN + KMPNN [23]	IND	ZINC15 (250K)	<1M	Link
	KV-PLM [129]	SMILES + Text	Transformer	MLM + MCM	PubChem(150M)+S2orc [63]	∼110 M	Link
	MOCO [135]	SMILES + FP + Graph + Geometry	Transformer + GIN + SchNet	IND	GEOM(50k)	—	—
	UniMol [132]	Geometry + Protein Pockets	Transformer	MCM + DM	ZINC/ChemBL + PDB [5]	—	Link
	MolT5 [21]	SMILES + Text	Transformer	Replace Corrupted Spans	C4 [75]+ ZINC-15 (100 M)	60M /770M	Link
	MICER [123]	SMILES + Image	CNNs + LSTM	AE	ZINC20	—	Link
	MM-Deacon [28]	SMILES + IUPAC	Transformer	IND	PubChem	10M	—
	PanGu Drug Model [58]	Graph + SELFIES [54]	Transformer + TransformerConv [91]	AE	ZINC20+ DrugSpaceX+UniChem(∼1.7 B)	∼104 M	Link
	ChemRL-GEM [22]	Graph + Geometry	GeoGNN [22]	MCM+CP	ZINC15 (20M)	—	Link

recover the masked tokens. Analogously, following the replace corrupted spans task of T5 [75], MolT5 [21] first masks some spans of abundant SMILES strings and biochemical text descriptions of molecules and then pre-train the Transformer model to predict the masked spans. In this way, these pre-trained models can generate both the SMILES strings and biochemical texts, which is especially effective for text-guided molecule generation and molecule caption (generation of the descriptive texts for molecules). Additionally, MICER [123] adopts an encoder–decoder based pre-training framework for molecular image caption (generation of the descriptive texts for molecular images). Specifically, they take molecular images as input to the pre-trained encoder (i.e. CNN), and then decode the corresponding SMILES strings. The above-mentioned multimodal pre-training strategies can advance the performance in the translations between various modalities. Additionally, various modalities can complement with each other and constitute a more complete knowledge base for downstream tasks [129].

5. Applications

The advancements in MPMs provide unprecedented opportunities to expedite a variety of end tasks involving molecular representations. In this section, we take drug discovery and development as an example and demonstrate several promising applications of MPMs in this domain. In Table 5, we summarize several widely-used datasets for evaluating MPMs in diverse applications.

5.1. Molecular Property Prediction (MPP)

In practice, bioactivities of a new drug candidate are controlled by many factors, such as solubility in the gastrointestinal tract, intestinal membrane permeability, and intestinal/hepatic first-pass metabolism [35]. However, wet-lab experiments are often laborious and expensive. As an alternative, MPMs can be utilized as a molecular encoder to obtain expressive representations for the new synthetic drug, which are helpful for downstream molecular property prediction tasks [80, 106].

MoleculeNet[1] [113] is the most common benchmark for molecular property prediction, which includes 700,000 molecules from PubChem [47], PubChem BioAssay [107], and ChEMBL [25]. The properties of molecules broadly fall into four categories: physiological, biophysical, physicochemical, and quantum mechanics. There are 17 datasets in MoleculeNet in total, among which FreeSolv, ESOL, MUV, HIV, BACE, BBBP, Tox21, ToxCast, SIDER, Clintox, QM7, QM8, and QM9 are the most commonly used ones to evaluate MPMs. The molecular property prediction using MoleculeNet can be regarded as multi-label binary classification or regression tasks in machine learning. More recently, a new quantum molecular dataset named Alchemy [8] is introduced for multi-task learning of molecular properties.

5.2. Molecular Generation (MG)

Molecular generation is a long-standing and promising research topic for computer-aided drug design. However, it is computationally prohibitive to enumerate the boundless drug-like chemical space [73]. Machine-learning-based approaches, especially generative models, have revolutionized the landscape of molecular generation via narrowing down the search space and improving computational efficiency [18, 136]. Currently, MPMs have already shown promising values in molecular generation. For example, MolGPT [4] adopts the autoregressive

[1] http://moleculenet.ai/

Table 5. Statistics of several widely-used datasets for downstream tasks involving molecular representations.

Dataset	Task	#Tasks	#Molecules	#Proteins	#Molecule–Protein	#Molecule–Molecule	Access Link
BBBP	MPP (Classification)	1	2,039	—	—	—	Link
Tox21	MPP (Classification)	12	7,831	—	—	—	Link
ToxCast	MPP (Classification)	617	8,576	—	—	—	Link
Sider	MPP (Classification)	27	1,427	—	—	—	Link
ClinTox	MPP (Classification)	2	1,478	—	—	—	Link
MUV	MPP (Classification)	17	93,087	—	—	—	Link
HIV	MPP (Classification)	1	41,127	—	—	—	Link
PCBA	MPP (Classification)	128	437,929	—	—	—	Link
Malaria	MPP (Regression)	1	9,999	—	—	—	Link
ESOL	MPP (Regression)	1	1,128	—	—	—	Link
FreeSolv	MPP (Regression)	1	643	—	—	—	Link
Lipophilicity	MPP (Regression)	1	4,200	—	—	—	Link
QM7	MPP (Regression)	1	6,830	—	—	—	Link
QM8	MPP (Regression)	12	21,786	—	—	—	Link
QM9	MPP (Regression)	3	133,885	—	—	—	Link
Alchemy	MPP (Regression)	12	119,487	—	—	—	Link
ZINC-250k	MG	1	250k	—	—	—	Link
ChEMBL	MG	1	2.1M	—	—	—	Link
TWOSIDES	DDI (Classification)	1	3,300	—	—	63,000	Link
DeepDDI	DDI (Classification)	1	192,284	—	—	19,187	Link
C. Elegans	DTI (Regression)	1	1,434	2,504	4,000 (positive interactions)	—	Link
Human	DTI (Regression)	1	1,502	852	3,369 (positive interactions)	—	Link

pre-training strategy to mimic the generation process of molecules, which shows promising results in terms of generating valid, unique, and novel molecules. Additionally, the emergence of multi-modal molecular pre-training techniques [21, 129] enables generation molecules from a descriptive text (text-to-molecule generation). In addition, molecular conformation generation which aims to generate 3D conformations of molecules has wide applications such as protein-ligand binding pose prediction [24, 121]. Traditional approaches based on molecular dynamics or Markov chain Monte Carlo are often computationally expensive, especially for large molecules [32]. The 3D-geometry-enhanced MPMs [132, 134] show notable superiority in the downstream task of conformation generation because they can capture the entailed relations between 2D molecular graph and 3D conformation. Representative datasets for the evaluation of molecule generation include ZINC [42], ChEMBL [2] and QM9 [76].

5.3. Drug-Drug Interaction (DDI)

Drug-drug interaction prediction is an imperative stage in drug discovery pipelines as it may lead to adverse drug reactions which will damage health or even cause death. Moreover, accurate DDI prediction can assist in medication recommendations. Therefore, DDI prediction serves as an indispensable part of the regulatory investigation before market approvals. From the machine learning aspect, DDI prediction can be regarded as a task that classifies the influence of combinative drugs into three categories: synergistic, additive, and antagonistic. Existing works of molecular pre-training, such as MPG [56], consider DDI prediction as a downstream task to evaluate the effectiveness of the MPMs. The commonly used datasets for DDI prediction include TWOSIDES[2] [98] and DeepDDI[3] [82] extracted from DrugBank [112].

5.4. Drug-Target Interaction (DTI)

Drug-target interaction prediction is a crucial task in the early stage of drug discovery that aims to identify drug candidates with the potency to bind to specific protein targets. Additionally, when a new disease appears, the best choice for therapy is to recycle approved drugs because of their availability and known safety profiles, i.e. drug repurposing [1]. Thus, DTI prediction can reduce the need for further drug discovery and lower drug safety risks. When leveraging MPMs for DTI, we need to both consider molecular encoders and also target encoders and predict binding affinities for DTI [70]. In this case, MPMs can be directly applied as a drug encoder and the pre-trained models provide good initializations for the molecule encoder. Then, the molecule encoder and target encoder are co-trained for the DTI prediction task. Existing works such as MPG [56] follow the aforementioned setting to advance DTI predictions. The widely-used datasets for DTI include Human and Caenorhabditis elegans[4].

[2] http://tatonettilab.org/offsides/

[3] https://bitbucket.org/kaistsystemsbiology/deepddi/src/master/

[4] http://admis.fudan.edu.cn/negative-cpi/

6. Conclusions and Future Outlooks

Despite the fruitful progress of MPMs, there are still major challenges that hamper the progress of the field. In this section, we highlight critical challenges and recommend promising research directions for future research.

6.1. Understanding Theoretical Groundings

Even though MPMs have demonstrated their capabilities in a variety of downstream tasks, rigorous theoretical analysis of MPMs has been lagging behind. It is vital for both the research and industry community to understand the key mechanism of what MPMs have learned and how it leads to improved performance for different downstream tasks. Then, we can better exploit the power of MPMs to amplify the strengths and avoid the weaknesses in real-world applications. For example, a recent study [96] observes that some self-supervised graph pre-training strategies do not always bring statistically significant advantages over non-pre-training counterparts. Further work is needed to fully understand the theoretical underpinnings of the success or failure of various molecular pre-training objectives.

6.2. Towards Better Knowledge Transfer

Tremendous efforts on MPMs have been devoted to pre-training strategies. However, it remains under-explored to understand how to leverage these pre-trained models to improve downstream task performance. Fine-tuning is a dominant technique to adapt the knowledge to various downstream tasks, whereas it is confronted with the issue of catastrophic forgetting [52], which means the MPMs often forget their learned knowledge during fine-tuning. To alleviate this issue, Han et al. [30] adaptively select and combine various pre-training tasks along with the target tasks in the fine-tuning stage to achieve a better adaptation. This strategy preserves sufficient knowledge captured by self-supervised pre-training tasks and improves the effectiveness of transfer learning on molecular pre-training. However, this strategy assumes that the pre-training tasks of MPMs are available for the downstream users, which is not practical in many real-world scenarios. Considerably more work will need to be done for better knowledge transfer from the pre-trained models.

6.3. Seeking Better Encoder Architectures and Pre-Training Tasks

As revealed in previous works [36, 38, 64], the applications of the powerful graph attention network (GAT) architecture for molecular graph pre-training will undermine the pre-training performance dramatically. It remains unknown why this phenomenon would occur and what kind of GNN architectures will be the most suitable alternative for molecular graph pre-training. On the other hand, for large-scale molecular graphs pre-training, how to integrate the message-passing scheme into Transformer as a unified encoder deserves more attention. Additionally, as we pointed out in Section 4, some representative pre-training strategies are still fraught with diverse issues. For example, MCM creates an undesirable gap between the pre-training and fine-tuning stages because it often contains artificial symbols that never occur in downstream tasks. More efforts are expected to mitigate these issues.

6.4. Seeking More Reliable and Realistic Benchmarks for Fair Evaluations

Current evaluation schemes of MPMs suffer from limited sizes, rendering the evaluation on these benchmarks less reliable for understanding the real progress of MPMs. Taking the popular quantum property prediction dataset QM9 as an example, existing MPMs have already achieved very high performance. Further competition and resources invested in those benchmarks may pose limited impacts on studying the problems. Another example is MoleculeNet which contains several expensive datasets for ADMET molecular property prediction [27]. However, since these datasets are quite small, the performance of the same model even varies by a large margin with different random seeds. Additionally, the other urgent need is to benchmark under a realistic setting, e.g., considering out-of-distribution generalization with scaffold splitting, where we split molecules according to the scaffold (molecular substructure). As in the real world, researchers always seek to apply MPMs from existing known molecules on unknown molecules, which may have different properties. Recently, Therapeutics Data Commons (TDC) is built to systematically access and evaluate machine learning models across a large number of therapeutic tasks [40], which could have the potential to serve as a unifying platform for a more fair evaluation of MPMs.

6.5. Pursuing Broader Impact in Diverse End Tasks

MPMs are often large in scale and consume much computing power to train. The ultimate goal is to obtain a universal molecular representation that can be utilized for any downstream tasks that involve molecules. However, critical gaps exist between methodology advancements and practical applications. On one side, MPMs have not been widely used to replace traditional molecular descriptors. On the other side, it remains under-explored how MPMs can benefit more downstream tasks including chemical reaction prediction [89], molecular similarity search in virtual screening [69], retrosynthesis [90], molecule design [18], chemical space exploration [19], etc.

Key Points

- *Structured taxonomy.* We contribute a structured taxonomy to provide a broad overview of the field, which categorizes existing works from four perspectives: molecular descriptors, neural architectures, pre-training strategies, and applications.
- *Current progress.* According to the taxonomy, we systematically delineate the current research directions on the topic of pre-trained models for molecules.
- *Abundant resources.* We collect abundant resources including open-sourced MPMs, available datasets, and an important paper list at https://github.com/junxia97/awesome-pretrain-on-molecules. We promise to continuously update these resources.

- *Future directions.* We discuss the limitations of existing works and suggest several promising future research directions.

Code Availability

The abundant resources including open-sourced MPMs, available datasets, downstream task datasets, and a paper list are available on GitHub: https://github.com/junxia97/awesome-pretrain-on-molecules.

Competing Interests

No competing interest is declared.

Short Descriptions of the Authors

Jun Xia is a Ph.D. candidate at Zhejiang University and Westlake University. His research interests focus on graph machine learning, drug discovery, and bioinformatics.
Yanqiao Zhu is a Ph.D. student at University of California at Los Angeles. His research interests lie primarily in the area of graph representation learning with an emphasis on self-supervised learning and interdisciplinary applications.
Yuanqi Du is a Ph.D. student at Cornell University. His research interests focus on deep generative models, geometric deep learning, and machine learning for scientific discovery.
Yue Liu is a visiting student at Westlake University. His research interests focus on graph machine learning and drug discovery.
Stan Z. Li received his Ph.D. degree from Surrey University, UK. He is currently a Chair Professor and director of the AI division of Westlake University. He was elevated to IEEE fellow and IAPR fellow for his contributions to the fields of face recognition, and artificial intelligence. His current research interests focus on machine learning and bioinformatics.

Author contribution statement

Jun Xia conceived this project. Yanqiao Zhu, Yuanqi Du, and Stan Z. Li wrote parts of this survey. Yue Liu collected available resources. All the authors revised and edited the manuscript.

Acknowledgments

We thank the editors and reviewers for their efforts in reviewing this manuscript. This work is supported in part by the Science and Technology Innovation 2030 – Major Project (No. 2021ZD0150100) and National Natural Science Foundation of China (No. U21A20427).

References

1. Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 2004.
2. AstraZeneca. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. *ChEMBL*, 2016.
3. Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, pages 1–10, 2021.
4. Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021.
5. Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.
6. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.
7. Deng Cai and Wai Lam. Graph transformer for graph-to-sequence learning. In *AAAI*, pages 7464–7471. AAAI Press, 2020.
8. Guangyong Chen, Pengfei Chen, Chang-Yu Hsieh, Chee-Kong Lee, Benben Liao, Renjie Liao, Weiwen Liu, Jiezhong Qiu, Qiming Sun, Jie Tang, et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv preprint arXiv:1906.09427*, 2019.
9. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.
10. Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.
11. Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*. OpenReview.net, 2020.
12. Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8):1757–1772, 2017.
13. Viviana Consonni and Roberto Todeschini. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References*. John Wiley & Sons, 2009.
14. Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Velickovic. Principal neighbourhood aggregation for graph nets. In *NeurIPS*, 2020.
15. Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3 (10):1503–1507, 2008.
16. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational

Linguistics, 2019.
17. Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. SE(3) equivariant graph neural networks with complete local frames. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 5583–5608. PMLR, 2022.
18. Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*, 2022.
19. Yuanqi Du, Xian Liu, Nilay Shah, Shengchao Liu, Jieyu Zhang, and Bolei Zhou. Chemspace: Interpretable and interactive chemical space exploration. 2022.
20. David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, pages 2224–2232, 2015.
21. Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.
22. Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.
23. Yin Fang, Qiang Zhang, and others. Molecular contrastive learning with chemical element knowledge graph. *AAAI*, 2022.
24. Octavian Ganea, Lagnajit Pattanaik, Connor W. Coley, Regina Barzilay, Klavs F. Jensen, William H. Green Jr., and Tommi S. Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. In *NeurIPS*, pages 13757–13769, 2021.
25. Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1): D1100–D1107, 2012.
26. Jonathan Godwin, Michael Schaarschmidt, Alexander L. Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Velickovic, James Kirkpatrick, and Peter W. Battaglia. Simple GNN regularisation for 3d molecular property prediction and beyond. In *ICLR*. OpenReview.net, 2022.
27. Joelle M. R. Gola, Olga Obrezanova, Ed Champness, and Matthew D. Segall. Admet property prediction: The state of the art and current challenges. *ChemInform*, 2006.
28. Zhihui Guo, Pramod Kumar Sharma, Andy Martinez, Liang Du, and Robin Abraham. Multilingual molecular representation learning via contrastive pre-training. In *ACL (1)*, pages 3441–3453. Association for Computational Linguistics, 2022.
29. William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pages 1024–1034, 2017.
30. Xueting Han, Zhenhuan Huang, Bang An, and Jing Bai. Adaptive transfer learning on graph neural networks. In *KDD*, pages 565–574. ACM, 2021.
31. Kaveh Hassani and Amir Hosein Khas Ahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 4116–4126. PMLR, 2020.
32. Paul CD Hawkins. Conformation generation: the state of the art. *Journal of chemical information and modeling*, 57(8):1747–1756, 2017.
33. R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*. OpenReview.net, 2019.
34. Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.
35. T. Hou, J. Wang, W. Zhang, and X. Xu. 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.
36. Yupeng Hou, Binbin Hu, Wayne Xin Zhao, Zhiqiang Zhang, Jun Zhou, and Ji-Rong Wen. Neural graph matching for pre-training graph neural networks. In *SDM*, pages 172–180. SIAM, 2022.
37. Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *KDD*, pages 594–604. ACM, 2022.
38. Weihua Hu, Bowen Liu, and others. Strategies for pre-training graph neural networks. In *ICLR*, 2020.
39. Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. GPT-GNN: generative pre-training of graph neural networks. In *KDD*, pages 1857–1867. ACM, 2020.
40. Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *NeurIPS Datasets and Benchmarks*, 2021.
41. Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Edge-augmented graph transformers: Global self-attention is enough for graphs. *arXiv preprint arXiv:2108.03348*, 2021.
42. John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012.
43. Stanisław Jastrzębski, Damian Leśniak, and Wojciech Marian Czarnecki. Learning to smile (s). *arXiv preprint arXiv:1602.06289*, 2016.
44. Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022.
45. José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.
46. Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.
47. Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He,

Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.
48. Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.
49. N. Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.
50. Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv:1611.07308*, 2016.
51. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.
52. James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
53. Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *ICLR*. OpenReview.net, 2020.
54. Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.
55. Pengyong Li, Jun Wang, Ziliang Li, Yixuan Qiao, Xianggen Liu, Fei Ma, Peng Gao, Sen Song, and Guotong Xie. Pairwise half-graph discrimination: A simple graph-level self-supervised strategy for pre-training graph neural networks. In *IJCAI*, pages 2694–2700. ijcai.org, 2021.
56. Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings Bioinform.*, 22(6), 2021.
57. Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *AAAI*, pages 4541–4549. AAAI Press, 2022.
58. Xinyuan Lin, Chi Xu, Zhaoping Xiong, Xinfeng Zhang, Ningxi Ni, Bolin Ni, Jianlong Chang, Ruiqing Pan, Zidong Wang, Fan Yu, et al. Pangu drug model: Learn a molecule like a human. *bioRxiv*, 2022.
59. Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*, 2022.
60. Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *ICLR*. OpenReview.net, 2022.
61. Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d molecular graphs. In *ICLR*, 2021.
62. Yixin Liu, Shirui Pan, and others. Graph self-supervised learning: A survey. *arXiv:2103.00111*, 2021.
63. Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*, 2019.
64. Yuanfu Lu, Xunqiang Jiang, Yuan Fang, and Chuan Shi. Learning to pre-train graph neural networks. In *AAAI*, pages 4276–4284. AAAI Press, 2021.
65. Nicholas A Meanwell. Synopsis of some recent tactical application of bioisosteres in drug design. *Journal of medicinal chemistry*, 2011.
66. Diego P. P. Mesquita, Amauri H. Souza Jr., and Samuel Kaski. Rethinking pooling in graph neural networks. In *NeurIPS*, 2020.
67. Grégoire Mialon, Dexiong Chen, Margot Selosse, and Julien Mairal. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:2106.05667*, 2021.
68. Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.
69. Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 2016.
70. Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.
71. Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C Stern, and Artem Cherkasov. The transformational role of gpu computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3):211–221, 2022.
72. Gabriel A Pinheiro, Juarez LF Da Silva, and Marcos G Quiles. Smiclr: Contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. *Journal of Chemical Information and Modeling*, 2022.
73. Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27(8):675–679, 2013.
74. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
75. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
76. Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):1–7, 2014.
77. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021.
78. Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. 2021.
79. David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
80. Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In

NeurIPS, 2020.

81. Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Molformer: Large scale chemical language representations capture molecular structure and properties. 2022.
82. Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 115(18):E4304–E4311, 2018.
83. Seongok Ryu, Jaechang Lim, Seung Hwan Hong, and Woo Youn Kim. Deeply learning molecular structure-property relationships using attention-and gate-augmented graph convolutional network. *arXiv preprint arXiv:1805.10988*, 2018.
84. Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 8459–8468. PMLR, 2020.
85. Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 2021.
86. Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. pages 991–1001, 2017.
87. Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1):1–8, 2017.
88. Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.
89. Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1):015016, 2021.
90. Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.
91. Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.
92. Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3d infomax improves gnns for molecular property prediction. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 20479–20502. PMLR, 2022.
93. Dagmar Stumpfe, Huabin Hu, and Juergen Bajorath. Evolving concept of activity cliffs. *ACS OMEGA*, pages 14360–14368, 2019.
94. Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*. OpenReview.net, 2020.
95. Mengying Sun, Jing Xing, and others. Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge. *KDD*, 2021.
96. Ruoxi Sun, Hanjun Dai, and Adams Wei Yu. Does GNN pretraining help molecular representation? In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=uytgM9N0vlR.
97. Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. Adversarial graph augmentation to improve graph contrastive learning. In *NeurIPS*, pages 15920–15933, 2021.
98. Nicholas P Tatonetti, Patrick P Ye, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125): 125ra31–125ra31, 2012.
99. Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Velickovic, and Michal Valko. Bootstrapped representation learning on graphs. *CoRR*, abs/2102.06514, 2021.
100. Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.
101. Thomas Unterthiner, Andreas Mayr, Günter Klambauer, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, and Sepp Hochreiter. Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS*, volume 27, pages 1–9, 2014.
102. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.
103. Petar Velickovic, Guillem Cucurull, and others. Graph attention networks. *ICLR*, 2018.
104. Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *ICLR (Poster)*. OpenReview.net, 2019.
105. Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436, 2019.
106. Y. Wang, J. Wang, and others. Molclr: Molecular contrastive learning of representations via graph neural networks. *ArXiv*, abs/2102.10056, 2021.
107. Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A Shoemaker, et al. Pubchem's bioassay database. *Nucleic acids research*, 40(D1):D400–D412, 2012.
108. Yanli Wang, Stephen H Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A Shoemaker, Paul A Thiessen, Siqian He, and Jian Zhang. Pubchem bioassay: 2017 update. *Nucleic acids research*, 45(D1):D955–D963, 2017.
109. Yingheng Wang, Yaosen Min, Erzhuo Shao, and Ji Wu. Molecular graph contrastive learning with parameterized explainable augmentations. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1558–1563. IEEE, 2021.
110. Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 2022.

111. David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
112. David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36 (suppl_1):D901–D906, 2008.
113. Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530, 2018.
114. Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z. Li. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *WWW*, pages 1070–1079. ACM, 2022.
115. Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, and Stan Z. Li. Progcl: Rethinking hard negative mining in graph contrastive learning. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 24332–24346. PMLR, 2022.
116. Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z. Li. Pre-training graph neural networks for molecular representations: Retrospect and prospect. In *ICML 2022 2nd AI for Science Workshop*, 2022. URL https://openreview.net/forum?id=dhXLkrY2Nj3.
117. Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.
118. Dongkuan Xu, Wei Cheng, Dongsheng Luo, Haifeng Chen, and Xiang Zhang. Infogcl: Information-aware graph contrastive learning. In *NeurIPS*, pages 30414–30425, 2021.
119. Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*. OpenReview.net, 2019.
120. Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 11548–11558. PMLR, 2021.
121. Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *ICLR*. OpenReview.net, 2022.
122. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764, 2019.
123. Jiacai Yi, Chengkun Wu, Xiaochen Zhang, Xinyi Xiao, Yanlong Qiu, Wentao Zhao, Tingjun Hou, and Dongsheng Cao. Micer: a pre-trained encoder–decoder architecture for molecular image captioning. *Bioinformatics*, 38(19): 4562–4572, 2022.
124. Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *NeurIPS*, pages 28877–28888, 2021.
125. Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.
126. Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 12121–12132. PMLR, 2021.
127. Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. Bringing your own view: Graph contrastive learning without prefabricated data augmentations. In *WSDM*, pages 1300–1309. ACM, 2022.
128. Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022.
129. Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):1–11, 2022.
130. Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S. Yu. From canonical correlation analysis to self-supervised graph neural networks. In *NeurIPS*, pages 76–89, 2021.
131. Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. In *NeurIPS*, pages 15870–15882, 2021.
132. Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-jjm0j-v2.
133. Jinhua Zhu, Yingce Xia, Tao Qin, Wen gang Zhou, Houqiang Li, and Tie-Yan Liu. Dual-view molecule pre-training. *ArXiv*, abs/2106.10234, 2021.
134. 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*, pages 2626–2636. ACM, 2022.
135. Yanqiao Zhu, Dingshuo Chen, Yuanqi Du, Yingze Wang, Qiang Liu, and Shu Wu. Improving molecular pretraining with complementary featurizations. *arXiv preprint arXiv:2209.15101*, 2022.
136. Yanqiao Zhu, Yuanqi Du, Yinkai Wang, Yichen Xu, Jieyu Zhang, Qiang Liu, and Shu Wu. A Survey on Deep Graph Generation: Methods and Applications. *arXiv.org*, March 2022.