A Systematic Survey of Chemical Pre-trained Models

Jun Xia[1] , **Yanqiao Zhu**[2] , **Yuanqi Du**[3] and **Stan Z. Li**[1]
[1]Research Center for Industries of the Future, Westlake University
[2]University of California, Los Angeles
[3]Cornell University
{xiajun, stan.zq.li}@westlake.edu.cn, yzhu@cs.ucla.edu, yd392@cs.cornell.edu

Abstract

Deep learning has achieved remarkable success in learning representations for molecules, which is crucial for various biochemical applications, ranging from property prediction to drug design. However, training Deep Neural Networks (DNNs) from scratch often requires abundant labeled molecules, which are expensive to acquire in the real world. To alleviate this issue, tremendous efforts have been devoted to Chemical Pre-trained Models (CPMs), where DNNs are pre-trained using large-scale unlabeled molecular databases and then fine-tuned over specific downstream tasks. Despite the prosperity, there lacks a systematic review of this fast-growing field. In this paper, we present the first survey that summarizes the current progress of CPMs. We first highlight the limitations of training molecular representation models from scratch to motivate CPM studies. Next, we systematically review recent advances on this topic from several key perspectives, including molecular descriptors, encoder architectures, pre-training strategies, and applications. We also highlight the challenges and promising avenues for future research, providing a useful resource for both machine learning and scientific communities.

1 Introduction

Extracting vector representations for molecules is critical to applying machine learning methods to a broad spectrum of molecular tasks. Initially, molecular fingerprints are developed to encode molecules into binary vectors with rule-based algorithms [Consonni and Todeschini, 2009]. Subsequently, various Deep Neural Networks (DNNs) have been employed to encode molecules in a data-driven manner. Early attempts exploit sequence-based neural architectures (e.g., RNNs, LSTMs, and transformers) to encode molecules represented in Simplified Molecular-Input Line-Entry System (SMILES) strings [Weininger, 1988]. 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 such structured inductive bias for better molecular representations [Kearnes *et al.*, 2016]. The key

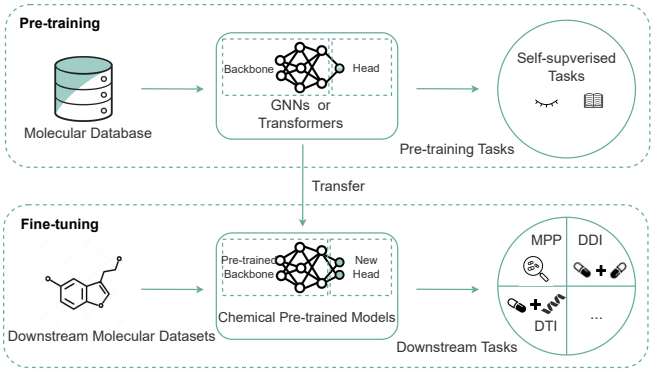


Figure 1: A typical learning pipeline for Chemical Pre-trained Models (CPMs). MPP: Molecular Property Prediction; DDI: Drug-Drug Interactions; DTI: Drug-Target Interactions.

advancements underneath these approaches are Graph Neural Networks (GNNs), which consider graph structures and attributive features simultaneously by recursively aggregating node features from neighborhoods [Kipf and Welling, 2017]. Recently, another line of development of GNNs for molecular representations models 3D geometric symmetries of molecular conformations, considering the molecules are in a constant motion in 3D space by nature [Schütt *et al.*, 2017]. However, the majority of the above works learn molecular representations under supervised settings, which limits their widespread application in practice for the following reasons. (1) *Scarcity of labeled data:* task-specific labels of molecules can be extremely scarce because molecular data labeling often requires expensive wet-lab experiments; (2) *Poor out-of-distribution generalization:* learning molecules with different sizes or functional groups requires out-of-distribution generalization in many real-world cases. For example, suppose one wishes to predict the properties of a newly synthesized molecule that differs from all the previous molecules in the training set. However, models trained from scratch cannot extrapolate to out-of-distribution molecules well [Hu *et al.*, 2020a].

Pre-trained Language Models (PLMs) have been a potential solution to the above challenges in the Natural Language Processing (NLP) community [Devlin *et al.*, 2019; Zheng *et al.*, 2022; Hu *et al.*, 2022]. Inspired by their success, as shown in Fig. 1, Chemical Pre-trained Models (CPMs)

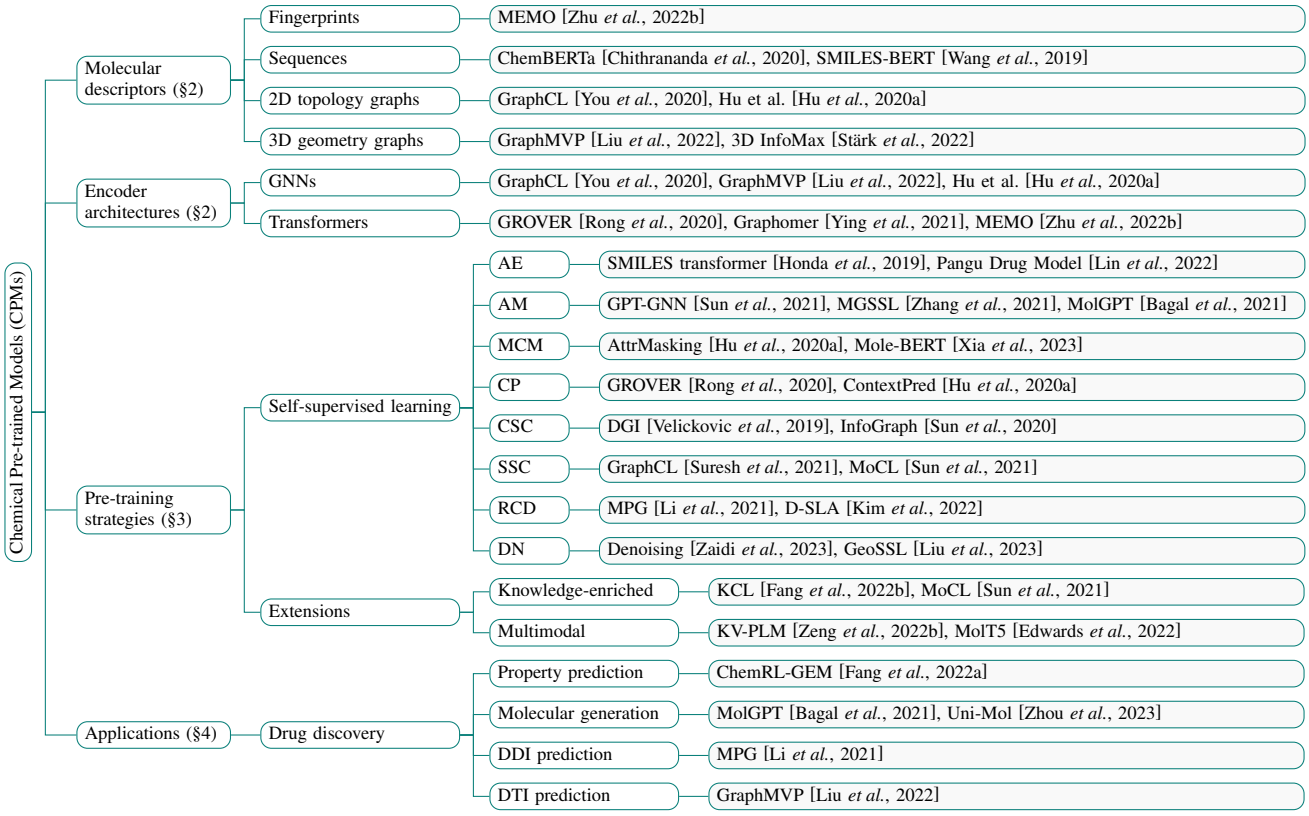


Figure 2: A taxonomy of Chemical Pre-trained Models (CPMs) with representative examples.

have been introduced to learn universal molecular representations from massive unlabeled molecules and then fine-tuned over specific downstream tasks. Initially, 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 [Devlin *et al.*, 2019]. This line of works include ChemBERTa [Chithrananda *et al.*, 2020], SMILES-BERT [Wang *et al.*, 2019], Molformer [Ross *et al.*, 2022], etc. More recently, the community explores pre-training on (both 2D and 3D) molecular graphs. For example, [Hu *et al.*, 2020a] propose to mask atom or edge attributes and predict the masked attributes. [Liu *et al.*, 2022] pre-train GNNs via maximizing the correspondence between 2D topological and 3D geometric structures.

Although CPMs have been increasingly applied in molecular representation learning, this rapidly expanding field still lacks a systematic review. In this paper, we present the first survey for CPMs to assist audiences of diverse backgrounds in understanding, using, and developing CPMs for various practical tasks. The contributions of this work can be summarized from the following four aspects. **(1)** *A structured taxonomy.* A broad overview of the field is presented with a structured taxonomy that categorizes existing works from four perspectives (Fig. 2): molecular descriptors, encoder architectures, pre-training strategies, and applications. **(2)** *Thorough review of the current progress.* Based on the taxonomy, the current research progress of pre-trained models for molecules is systematically delineated. **(3)** *Abundant additional resources.* Abundant resources including open-sourced CPMs, available datasets, and an important paper list are collected and can be found at https://github.com/junxia97/awesome-pretrain-on-molecules. These resources will be continuously updated on a regular basis. **(4)** *Discussion of future directions.* The limitations of existing works are discussed and several promising research directions are highlighted.

2 Molecular Descriptors and Encoders

In order to feed molecules to DNNs, molecules have to be featurized in numerical descriptors. Various descriptors are designed to describe molecules in a concise format. In this section, we briefly review these molecular descriptors and their corresponding neural encoder architectures.

Fingerprints (FP). Molecular fingerprints describe the presence or absence of particular substructures of a molecule with binary strings. For example, PubChemFP [Wang *et al.*, 2017] encodes 881 structural key types that correspond to the substructures for a fragment of compounds in the PubChem database.

Sequences. The most frequently used sequential descriptor for molecules is the Simplified Molecular-Input Line-Entry System (SMILES) [Weininger, 1988] 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. Transformers [Vaswani

et al., 2017] are a powerful neural model for processing sequences and modeling the complex relationships among each token. We can split sequence-based molecular descriptors into a series of tokens denoting atoms/bonds at first and then apply transformers on top of these tokens [Chithrananda *et al.*, 2020; Wang *et al.*, 2019].

2D graphs. Molecules can be represented as 2D graphs naturally, with atoms as nodes and bonds as edges. Each node and edge can also carry feature vectors denoting the atom types/chirality and bond types/direction for instance [Hu *et al.*, 2020a]. Here, GNNs [Kipf and Welling, 2017; Xu *et al.*, 2019] can be used to learn 2D molecular graph representations. Some hybrid architectures of GNNs and transformers [Rong *et al.*, 2020; Ying *et al.*, 2021] can also be leveraged to capture the topological structures of molecular graphs.

3D graphs. 3D geometries represent the spatial arrangements of atoms of the molecule in the 3D space, where each atom is associated with its type and coordinate plus some optional geometric attributes such as velocity. The advantage of using 3D geometry is that the conformer information is critical to many molecular properties, especially quantum properties. In addition, it is also possible to directly leverage stereochemistry information such as chirality given the 3D geometries. A number of approaches [Schütt *et al.*, 2017; Satorras *et al.*, 2021; Du *et al.*, 2022a] have developed message-passing mechanisms on 3D geometries, which enable the graph representations to follow certain physical symmetries, such as equivariance to translations and rotations.

3 Pre-training Strategies

In this section, we elaborate on several representative self-supervised pre-training strategies of CPMs.

3.1 AutoEncoding (AE)

Reconstructing molecules with autoencoders (Fig. 3a) serves as a natural self-supervised target for learning expressive molecular representations. The prediction in molecule reconstructions is (partial) structures of the given molecules such as the attributes of a subset of atoms or chemical bonds. A typical example is SMILES transformer [Honda *et al.*, 2019], which leverages a transformer-based encoder-decoder network and learns the representations by reconstructing molecules represented by SMILES strings. More recently, unlike conventional autoencoders with the same types for the input and output data, [Lin *et al.*, 2022] pre-train a graph-to-sequence asymmetric conditional variational autoencoder to learn molecular representations. Although autoencoders can learn meaningful representations for molecules, they focus on single molecules and fail to capture inter-molecule relationships, which limits their performance in some downstream tasks [Li *et al.*, 2021].

3.2 Autoregressive Modeling (AM)

Autoregressive Modeling (AM) factorizes the molecular input as a sequence of sub-components and then it predicts the sub-components one by one, conditioned on previous sub-components in the sequence. Following the idea of GPT [Brown *et al.*, 2020] in NLP, MolGPT [Bagal *et al.*, 2021] pre-trains a transformer network to predict the next token in the SMILES strings in such an autoregressive manner. For molecular graphs, GPT-GNN [Hu *et al.*, 2020b] 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. Analogously, MGSSL [Zhang *et al.*, 2021] generates molecular graph motifs instead of individual atoms or bonds autoregressively. Formally, such autoregressive modeling objectives can be written as

$$\mathcal{L}_{\text{AM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \sum_{i=1}^{|\mathcal{C}|} \log p(\mathcal{C}_i \mid \mathcal{C}_{<i}), \tag{1}$$

where $\mathcal{C}_i, \mathcal{C}_{<i}$ are the attributes of i-th component and the ones generated before the index i in the molecule $\mathcal{M}$, respectively. Compared with other strategies, AM allows CPMs to perform better at generating molecules its training procedure resembles that of molecule generation [Bagal *et al.*, 2021]. However, AM is more computationally expensive and requires a preset ordering of atoms or bonds beforehand, which may be inappropriate for molecules because the atoms or bonds do not present inherent orders.

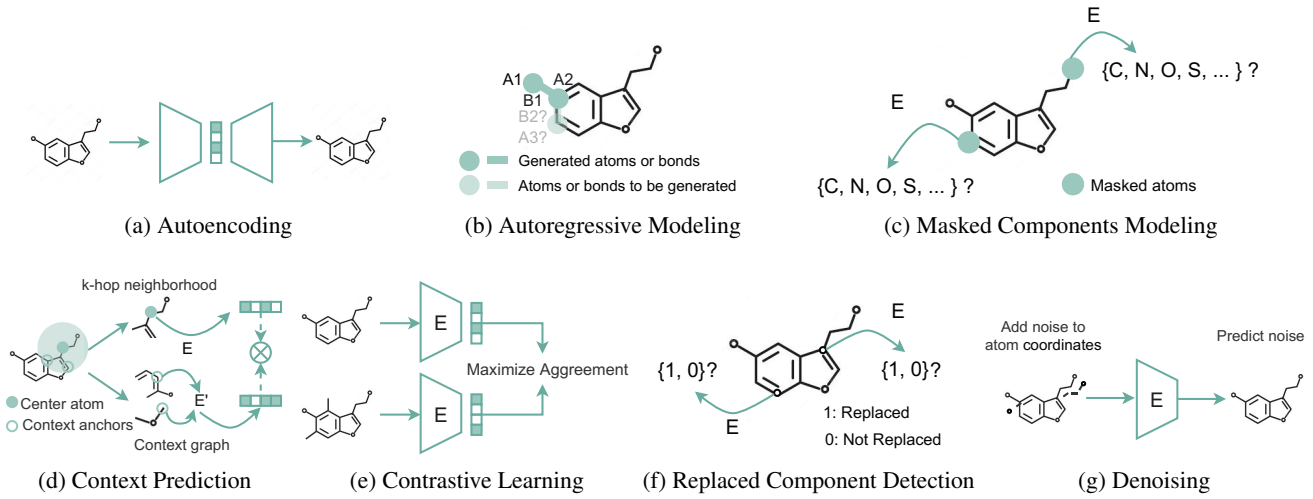
3.3 Masked Component Modeling (MCM)

In language domains, Masked Language Modeling (MLM) has emerged as a dominant pre-training objective. Specifically, MLM randomly masks out tokens from the input sentences, where the model can be trained to predict those masked tokens using the remaining tokens [Devlin *et al.*, 2019]. Masked Component Modeling (MCM, Fig. 3c) generalizes the idea of MLM for molecules. Specifically, MCM masks out some components (e.g., atoms, bonds, and fragments) of the molecules and then trains the model to predict them given the remaining components. Generally, its objective can be formulated as

$$\mathcal{L}_{\text{MCM}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \sum_{\widetilde{\mathcal{M}}\in m(\mathcal{M})} \log p(\widetilde{\mathcal{M}} \mid \mathcal{M}\backslash m(\mathcal{M})), \tag{2}$$

where $m(\mathcal{M})$ denotes the masked components from the molecule $\mathcal{M}$ and $\mathcal{M}\backslash m(\mathcal{M})$ are the remaining components. For sequence-based pre-training, ChemBERTa [Chithrananda *et al.*, 2020], SMILES-BERT [Wang *et al.*, 2019], and Molformer [Ross *et al.*, 2022] mask random characters in the SMILES strings and then recover them based on the output of the transformer of the corrupted SMILES strings. For molecular graph pre-training, [Hu *et al.*, 2020a] propose to randomly mask input atom/chemical bond attributes and pre-train the GNNs to predict them. Similarly, GROVER [Rong *et al.*, 2020] attempts to predict the masked subgraphs to capture the contextual information in the molecular graphs. Recently, Mole-BERT [Xia *et al.*, 2023] argues that masking the atom types could be problematic due to the extremely small and unbalanced atom set in nature. To mitigate this issue, they design a context-aware tokenizer to encode atoms as chemically meaningful discrete values for masking.

MCM is especially beneficial for richly-annotated molecules. For example, masking atom attributes enables



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

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

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

GNNs to learn simple chemistry rules such as valency, as well as potentially other complex chemistry descriptors such as the electronic or steric effects of functional groups. Additionally, compared with the aforementioned AM strategies, MCM predicts the masked components based on their surrounding environments as opposed to AM which merely relies on preceding components in a predefined sequence. Hence, MCM can capture more complete chemical semantics. However, as MCM often masks a fixed portion of each molecule during pre-training following BERT [Devlin *et al.*, 2019], it cannot train on all the components in each molecule, which results in less efficient sample utilization.

3.4 Context Prediction (CP)

Context Prediction (CP, Fig. 3d) aims to capture the semantics of molecules/atoms in an explicit, context-aware manner. Generally, CP can be formulated as

$$\mathcal{L}_{\text{CP}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}} \log p\left(t \mid \mathcal{M}_1, \mathcal{M}_2\right), \tag{3}$$

where $t = 1$ if neighborhood components $\mathcal{M}_1$ and surrounding contexts $\mathcal{M}_2$ share the same center atom and otherwise $t = 0$. For example, [Hu *et al.*, 2020a] use a binary classification of whether the subgraphs in molecules and surrounding context structures belong to the same node. While simple and effective, CP requires an auxiliary neural model to encode the context into a fixed vector, adding extra computational overhead for large-scale pre-training.

3.5 Contrastive Learning (CL)

Contrastive Learning (CL, Fig. 3e) pre-trains the model by maximizing the agreement between a pair of similar inputs, such as two different augmentations or descriptors of the same molecule. According to the contrastive granularity (e.g., molecule- or substructure-level), we introduce two categories of CL in CPMs: Cross-Scale Contrast (CSC) and Same-Scale Contrast (SSC).

Cross-Scale Contrast (CSC). Deep InfoMax is a representative CSC model that is originally proposed for learning image representations by contrasting a pair of an image and its local regions against other negative pairs [Hjelm *et al.*, 2019]. For molecular graphs, InfoGraph [Sun *et al.*, 2020] follows this idea by contrasting molecule- and substructure-level representations, which can be formally described as

$$\mathcal{L}_{\text{CSC}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}\left[\log s(\mathcal{M},\mathcal{C}) - \log \sum_{\mathcal{C}^-\in\mathcal{N}} s(\mathcal{M},\mathcal{C}^-)\right], \tag{4}$$

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 [Hassani and Ahmadi, 2020] performs node diffusion to generate an augmented molecular graph and then maximizes the similarity between original and augmented views by contrasting atom representations of one view with molecular representations of the other view and vice versa.

Same-Scale Contrast (SSC). Same-Scale Contrast (SSC) performs contrastive learning on individual molecules by pushing the augmented molecule close to the anchor molecule (positive pairs) and away from other molecules (negative pairs). For example, GraphCL [You *et al.*, 2020] and its variants [You *et al.*, 2021; Sun *et al.*, 2021; Suresh *et al.*, 2021; Xu *et al.*, 2021a; Fang *et al.*, 2022b; Wang *et al.*, 2021; Xia *et al.*, 2022b; Wang *et al.*, 2022a] propose various augmentation strategies for molecule-level pre-training represented by graphs. Additionally, some recent works maximize the agreement between various descriptors of identical molecules and repel the different ones. For example, SMICLR [Pinheiro *et al.*, 2022] jointly leverages a graph encoder and a SMILES string encoder to perform SSC; MM-Deacon [Guo *et al.*, 2022] utilizes two separate transformers to encode the SMILES and the International Union of Pure and Applied Chemistry (IUPAC) of molecules, after which a contrastive objective is used to promote similarity of SMILES and IUPAC representations

from the same molecule; 3DInfoMax [Stärk *et al.*, 2022] proposes to maximize the agreements between the learned 3D geometry and 2D graph representations; GeomGCL [Li *et al.*, 2022b] adopts a dual-view Geometric Message Passing Neural Network (GeomMPNN) to encode both 2D and 3D graphs of a molecule and design a geometric contrastive objective. The general formulation of the SSC pre-training objective is

$$\mathcal{L}_{\text{SSC}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}\left[\log s(\mathcal{M},\mathcal{M}') - \log \sum_{\mathcal{M}^-\in\mathcal{N}} s(\mathcal{M},\mathcal{M}^-)\right], \tag{5}$$

where $\mathcal{M}'$ is the augmented versions or other descriptors of the molecule $\mathcal{M}$, and $\mathcal{N}$ is the set of negative samples.

Although CL has achieved promising results, several critical issues impede its broader applications. Firstly, it is difficult to preserve semantics during molecular augmentations. Existing solutions pick augmentations with manual trial-and-errors [You *et al.*, 2020], cumbersome optimization [You *et al.*, 2021], or through the guidance of expensive domain knowledge [Sun *et al.*, 2021], but an efficient and principled way to design the chemically appropriate augmentation for molecular pre-training is still lacking. Also, the assumption behind CL that pulls similar representations closer may not always hold true for molecular representation learning. For example, in the case of molecular activity cliffs [Stumpfe *et al.*, 2019], similar molecules hold completely different properties. Therefore, it remains unsolved which pre-training strategies can better capture discrepancies between molecules. Additionally, the CL objective in most CPMs randomly chooses all other molecules in one batch as the negative samples regardless of their true semantics, which will undesirably repel the molecules of similar properties and undermine the performance due to the false negatives [Xia *et al.*, 2022a].

3.6 Replaced Components Detection (RCD)

Replaced Components Detection (RCD, Fig. 3f) proposes to recognize randomly replaced components of the input molecules. For example, MPG [Li *et al.*, 2021] splits each molecule into two parts, shuffles their structure by combining parts from two molecules, and trains the encoder to detect whether the combined parts belong to the same molecule. This objective can be written as

$$\mathcal{L}_{\text{RCD}} = -\mathbb{E}_{\mathcal{M}\in\mathcal{D}}\left[\log p(t \mid \mathcal{M}_1, \mathcal{M}_2)\right], \tag{6}$$

where $t = 1$ if the two parts $\mathcal{M}_1$ and $\mathcal{M}_2$ are from the same molecule $\mathcal{M}$ and $t = 0$ otherwise. While RCD can uncover intrinsic patterns in molecular structures, the encoders are pre-trained to always produce the same "non-replacement" label for all natural molecules and "replacement" label for randomly combined molecules. However, in downstream tasks, the input molecules are all natural ones, causing the molecular representations produced by RCD to be less distinguishable.

3.7 DeNoising (DN)

Inspired by the success of denoising diffusion probabilistic models [Ho *et al.*, 2020], DeNoising (DN, Fig. 3g) has been recently adopted as a pre-training strategy for learning molecular representations as well. A recent work [Zaidi *et al.*, 2023] adds noise to atomic coordinates of 3D molecular geometry and pre-trains the encoders to predict the noise. They demonstrate that such a denoising objective approximates learning a molecular force field. A concurrent work, Uni-Mol [Zhou *et al.*, 2023] adds noise to atomic coordinates motivated by the fact that masked atom types can be easily inferred given 3D atomic positions. More recently, GeoSSL [Liu *et al.*, 2023] proposes a distance denoising pre-training method to model the dynamic nature of 3D molecules. Generally, the pre-training objective of denoising can be formulated as

$$\mathcal{L}_{\text{DN}} = \mathbb{E}_{\mathcal{M}\in\mathcal{D}}\|\epsilon - f_\theta(\widetilde{\mathcal{M}})\|^2, \tag{7}$$

where ϵ denotes the added noise, $\widetilde{\mathcal{M}}$ denotes the input molecule $\mathcal{M}$ with noise added, and $f_\theta(\cdot)$ denotes the encoders that predict the noise.

3.8 Extensions

Knowledge-enriched pre-training. CPMs usually learn general molecular representations from a large molecular database. However, they often lack domain-specific knowledge. To improve their performance, several recent works try to inject external knowledge into CPMs. For example, GraphCL [You *et al.*, 2020] first points out that bond perturbations (adding or dropping the bonds as data augmentations) are conceptually incompatible with domain knowledge and empirically not helpful for contrastive pre-training on chemical compounds. Therefore, they avoid adopting bond perturbations for molecular graph augmentation. More explicitly, MoCL [Sun *et al.*, 2021] proposes a domain knowledge-based molecular augmentation operator called substructure substitution, in which a valid substructure of a molecule is replaced by a bioisostere which produces a new molecule with similar physical or chemical properties as the original one. More recently, KCL [Fang *et al.*, 2022b] constructs a chemical element Knowledge Graph (KG) to summarize microscopic associations between chemical elements and presents a novel Knowledge-enhanced Contrastive Learning (KCL) framework for molecular representation learning. Additionally, MGSSL [Zhang *et al.*, 2021] first leverages existing algorithms [Degen *et al.*, 2008] to extract semantically meaningful motifs and then pre-trains neural encoders to predict the motifs in an autoregressive manner. ChemRL-GEM [Fang *et al.*, 2022a] proposes to utilize molecular geometry information to enhance molecular graph pre-training. It designs a geometry-based GNN 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. Although knowledge-enriched pre-training helps CPMs capture chemical domain knowledge, it requires expensive prior knowledge as guidance, which poses a hurdle to broader applications when the prior is incomplete, incorrect, or expensive to obtain.

Multimodal pre-training. In addition to the descriptors mentioned in Sec. 2, molecules can also be described using other modalities including images and biochemical texts. Some recent works perform multimodal pre-training on molecules. For example, KV-PLM [Zeng *et al.*, 2022b] first

	Model	Input	Backbone architecture	Pre-training task	Pre-training database	#Params.	Link
Sequence	SMILES Transformer [Honda *et al.*, 2019]	SMILES	Transformer	AE	ChEMBL (861K) [Gaulton *et al.*, 2012]	—	Link
	ChemBERTa [Chithrananda *et al.*, 2020]	SMILES/SELFIES [Krenn *et al.*, 2020]	Transformer	MCM	PubChem (77M) [Wang *et al.*, 2017]	—	Link
	SMILES-BERT [Wang *et al.*, 2019]	SMILES	Transformer	MCM	ZINC15 (∼18.6M) [Sterling and Irwin, 2015]	—	Link
	Molformer [Ross *et al.*, 2022]	SMILES	Transformer	MCM	ZINC15 (1B) + PubChem (111M)	—	—
Graph / Geometry	Hu et al. [Hu *et al.*, 2020a]	Graph	5-layer GIN	CP + MCM	ZINC15 (2M) + ChEMBL (456K)	∼2M	Link
	GraphCL [You *et al.*, 2020]	Graph	5-layer GIN	SSC	ZINC15 (2M)	∼2M	Link
	JOAO [You *et al.*, 2021]	Graph	5-layer GIN	SSC	ZINC15 (2M)	∼2M	Link
	AD-GCL [Suresh *et al.*, 2021]	Graph	5-layer GIN	SSC	ZINC15 (2M)	∼2M	Link
	GraphLoG [Xu *et al.*, 2021b]	Graph	5-layer GIN	SSC	ZINC15 (2M)	∼2M	Link
	MGSSL [Zhang *et al.*, 2021]	Graph	5-layer GIN	MCM + AM	ZINC15 (250K)	∼2M	Link
	MPG [Li *et al.*, 2021]	Graph	MolGNet [Li *et al.*, 2021]	RCD + MCM	ZINC + ChEMBL (11M)	53M	Link
	LP-Info [You *et al.*, 2022]	Graph	5-layer GIN	SSC	ZINC15 (2M)	∼2M	Link
	SimGRACE [Xia *et al.*, 2022b]	Graph	5-layer GIN	SSC	ZINC15 (2M)	∼2M	Link
	GraphMAE [Hou *et al.*, 2022]	Graph	5-layer GIN	AE	ZINC15 (2M)	∼2M	Link
	MGMAE [Feng *et al.*, 2022]	Graph	5-layer GIN	AE	ZINC15 (2M) + ChEMBL (456K)	∼2M	—
	GROVER [Rong *et al.*, 2020]	Graph	GTransformer [Rong *et al.*, 2020]	CP + MCM	ZINC + ChEMBL (10M)	48M∼100M	Link
	MolCLR [Wang *et al.*, 2022b]	Graph	GCN + GIN	SSC	PubChem (10M)	—	Link
	Graphomer [Ying *et al.*, 2021]	Graph	Graphomer [Ying *et al.*, 2021]	Supervised	PCQM4M-LSC (∼3.8M) [Hu *et al.*, 2021]	—	Link
	3D-EMGP [Jiao *et al.*, 2023]	Geometry	E(3)-equivariant GNNs	DN	GEOM (100K) [Axelrod and Gómez-Bombarelli, 2022]	—	Link
	Mole-BERT [Xia *et al.*, 2023]	Graph	5-layer GIN	MCM + SSC	ZINC15 (2M)	∼2M	Link
	Denoising [Zaidi *et al.*, 2023]	Geometry	GNS [Sanchez-Gonzalez *et al.*, 2020]	DN	PCQM4Mv2 (∼3.4M)	—	Link
	GeoSSL [Liu *et al.*, 2023]	Geometry	PaiNN [Schütt *et al.*, 2021]	DN	Molecule3D [Xu *et al.*, 2021c] (∼1M)	—	Link
Multimodal / External knowledge	DMP [Zhu *et al.*, 2021]	Graph + SMILES	DeeperGCN + Transformer	MCM + SSC	PubChem (110M)	104.1M	Link
	GraphMVP [Liu *et al.*, 2022]	Graph + Geometry	5-layer GIN + SchNet [Schütt *et al.*, 2017]	SSC + AE	GEOM (50K)	∼2M	Link
	3D Infomax [Stärk *et al.*, 2022]	Graph + Geometry	PNA [Corso *et al.*, 2020]	SSC	QM9 (50K) + GEOM (140K) + QMugs (620K)	—	Link
	KCL [Fang *et al.*, 2022b]	Graph + Knowledge Graph	GCN + KMPNN [Fang *et al.*, 2022b]	SSC	ZINC15 (250K)	<1M	Link
	KV-PLM [Zeng *et al.*, 2022b]	SMILES + Text	Transformer	MLM + MCM	PubChem (150M)	∼110M	Link
	MEMO [Zhu *et al.*, 2022b]	SMILES + FP + Graph + Geometry	Transformer + GIN + SchNet	SSC	GEOM (50K)	—	—
	MolT5 [Edwards *et al.*, 2022]	SMILES + Text	Transformer	Replace Corrupted Spans	ZINC-15 (100M)	60M / 770M	Link
	MICER [Yi *et al.*, 2022]	SMILES + Image	CNNs + LSTM	AE	ZINC20	—	Link
	MM-Deacon [Guo *et al.*, 2022]	SMILES + IUPAC	Transformer	SSC	PubChem	10M	—
	PanGu Drug Model [Lin *et al.*, 2022]	Graph + SELFIES [Krenn *et al.*, 2020]	Transformer	AE	ZINC20 + DrugSpaceX + UniChem (∼1.7B)	∼104M	Link
	KPGT [Li *et al.*, 2022a]	SMILES + FP	LiGhT [Li *et al.*, 2022a]	MCM	ChEMBL29 (2M)	—	Link
	ChemRL-GEM [Fang *et al.*, 2022a]	Graph + Geometry	GeoGNN [Fang *et al.*, 2022a]	MCM+CP	ZINC15 (20M)	—	Link
	ImageMol [Zeng *et al.*, 2022a]	Molecular Images	ResNet18 [He *et al.*, 2016]	AE + SSC + CP	PubChem (∼10M)	—	Link
	Uni-Mol [Zhou *et al.*, 2023]	Geometry + Protein Pockets	Transformer	MCM + DN	ZINC/ChemBL + PDB [Berman *et al.*, 2000]	—	Link

Table 1: A summary of representative Chemical Pre-trained Models (CPMs) in literature.

tokenizes both the SMILES strings and biochemical texts. Then, they randomly mask part of the tokens and pre-train the neural encoders to recover the masked tokens. Analogously, following the replace corrupted spans task of T5 [Raffel *et al.*, 2020], MolT5 [Edwards *et al.*, 2022] first masks some spans of abundant SMILES strings and biochemical text descriptions of molecules and then pre-train the transformer 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 captioning (generation of the descriptive texts for molecules). [Zhu *et al.*, 2022b] propose to maximize the consistency between the embeddings of four molecular descriptors and their aggregated embedding using a contrastive objective. In this way, these various descriptors can collaborate with each other for molecular property prediction tasks. Additionally, MICER [Yi *et al.*, 2022] adopts an autoencoder-based pre-training framework for molecular image captioning. Specifically, they feed molecular images to the pre-trained encoder and then decode the corresponding SMILES strings. The above-mentioned multimodal pre-training strategies can advance the translations between various modalities. Also, these modalities can work together to create a more complete knowledge base for various downstream tasks.

4 Applications

The following section takes drug discovery as a case study and showcases several promising applications of CPMs (Tab. 1).

4.1 Molecular Property Prediction (MPP)

The bioactivity of a new drug candidate is influenced by various factors in real life, including solubility in the gastrointestinal tract, intestinal membrane permeability, and intestinal/hepatic first-pass metabolism. However, such labels for molecules can be extremely scarce because wet-lab experiments are often laborious and expensive. CPMs offer a solution that can exploit the massive unlabeled molecules and serve as powerful backbones for downstream molecular property prediction tasks [Wang *et al.*, 2022b; Zaidi *et al.*, 2023]. Furthermore, compared with the models trained from scratch, CPMs can better extrapolate to out-of-distribution molecules, which is especially important when predicting the properties of newly synthesized drugs [Hu *et al.*, 2020a].

4.2 Molecular Generation (MG)

Molecular generation, a long-standing challenge in computer-aided drug design, has been revolutionized by machine learning methods, especially generative models, that narrow the search space and improve computational efficiency, making it possible to delve into the seemingly infinite drug-like chemical space [Du *et al.*, 2022b]. CPMs, such as MolGPT [Bagal *et al.*, 2021], which employs an autoregressive pre-training approach, has proven to be instrumental in generating valid, unique, and novel molecular structures. The emergence of multi-modal molecular pre-training techniques [Edwards *et al.*, 2022; Zeng *et al.*, 2022b] has further expanded the possibilities of molecular generation by enabling the transformation of descriptive text into molecular structures. Another crucial area where CPMs have demonstrated their prowess is the generation of three-dimensional molecular conformations, particularly for the prediction of protein-ligand binding poses. Unlike conventional approaches based on molecular dynamics or Markov chain Monte Carlo, which are often hindered by computational limitations, especially for larger molecules [Hawkins, 2017], CPMs based on 3D geometry [Zhu *et al.*, 2022a; Zhou *et al.*, 2023] exhibit remarkable superiority in conformation generation tasks, as they can capture some inherent relationships between 2D molecules and 3D conformations during the pre-training process.

4.3 Drug-Target Interactions (DTI)

Predictive analysis of Drug-Target Interactions (DTI) is a vital step in the early stages of drug discovery, as it helps to identify drug candidates with binding potential to specific protein targets. This is particularly important in drug repurposing, where the goal is to recycle approved drugs for a new disease, thereby reducing the need for further drug discovery and minimizing safety risks. Also, obtaining sufficient drug-target data for supervised training can be challenging. CPMs can overcome this issue by providing molecular encoders with good initializations. To achieve accurate DTI predictions, it is essential to consider both molecular encoders and target encoders, predict binding affinities, and co-train both for the DTI prediction task [Nguyen *et al.*, 2021]. Previous works such as MPG [Li *et al.*, 2021] follow these principles to advance DTI predictions.

4.4 Drug-Drug Interactions (DDI)

Accurately predicting Drug-Drug Interactions (DDI) is another crucial stage in drug discovery pipelines as such interactions can result in adverse reactions that can harm health and even cause death. Moreover, accurate DDI predictions can also assist in making informed medication recommendations, making it an essential part of the regulatory investigation prior to market approval. From the machine learning perspective, DDI prediction can be regarded as a classification task that determines the influence of combination drugs as synergistic, additive, or antagonistic. To achieve effective DDI prediction, expressive molecular representations are required, which can be obtained using CPMs. MPG [Li *et al.*, 2021] is a representative example that has demonstrated the usefulness of CPMs by adopting DDI prediction as a downstream task.

5 Conclusions and Future Outlooks

In conclusion, this paper provides a comprehensive overview of Chemical Pre-trained Models. We start by reviewing the widely-used descriptors and encoders for molecules, then present the representative pre-training strategies and evaluate their advantages and disadvantages. We also showcase various successful applications of CPMs in drug discovery and development. Despite the fruitful progress, there are still several challenges that warrant further research in the future.

5.1 Improving Pre-training Architectures

While remarkable advancements have been achieved in analyzing the learning capabilities of neural architectures, such as the WL-test for GNNs, these analyses lack specificity in determining the optimal design for highly structured molecules. The ideal featurizations and architectures for CPMs remain elusive, as evidenced by conflicting results such as the negative impact of Graph Attention Networks (GATs) [Velickovic *et al.*, 2018], which are widely adopted in graph learning, on downstream performance in previous studies [Hu *et al.*, 2020a; Hou *et al.*, 2022]. Furthermore, there is a pressing need to explore ways of seamlessly integrating message-passing techniques into transformers as a unified encoder to accommodate pre-training of large-scale molecular graphs. Additionally, as discussed in Sec. 3, the pre-training objectives still leave much room for improvement, with the efficient masking strategy of subcomponents in MCM being a prime example.

5.2 Building Reliable and Realistic Benchmarks

Despite the numerous studies conducted on CPMs, their experimental results can sometimes be unreliable due to the inconsistent evaluation settings employed (e.g., random seeds and dataset splits). For instance, on MoleculeNet [Wu *et al.*, 2018] that contains several expensive datasets for molecular property prediction, the performance of the same model can vary significantly with different random seeds, possibly due to the relatively small scale of these molecular datasets. It is also crucial to establish more reliable and realistic benchmarks for CPMs that take out-of-distribution generalization into account. One solution is to evaluate CPMs through scaffold splitting, which involves splitting molecules based on their substructures. In reality, researchers must often apply CPMs trained from already known molecules to newly synthesized, unknown molecules that may differ greatly in properties and belong to divergent domains. In this regard, the recently established Therapeutics Data Commons (TDC) [Huang *et al.*, 2021] offers a promising opportunity to fairly evaluate CPMs across a diverse range of therapeutic applications.

5.3 Broadening the Impact of CPMs

The ultimate goal of CPMs is to develop versatile molecular encoders that can be applied to various downstream tasks related to molecules. Compared to the progress of PLMs in the NLP community, there remains a substantial disparity between methodological advancements and practical applications of CPMs. On one hand, the representations produced by CPMs have not yet been extensively used to replace conventional molecular descriptors in chemistry, and these pre-trained models have not yet become standard tools for the community. On the other hand, there is limited exploration into how these models can benefit a more extensive range of downstream tasks beyond individual molecules, such as chemical reaction prediction, molecular similarity searches in virtual screening, retrosynthesis, chemical space exploration, and many others.

5.4 Establishing Theoretical Foundations

Even though CPMs have demonstrated impressive performance in various downstream tasks, a rigorous theoretical understanding of these models has been limited. This lack of underpinnings presents a hindrance to both the scientific community and industry stakeholders who seek to maximize the potential of these models. The theoretical foundations of CPMs must be established in order to fully comprehend their mechanisms and how they drive improved performance in various applications. For example, a recent empirical study [Sun *et al.*, 2022] has questioned the superiority of certain self-supervised graph pre-training strategies over non-pre-trained counterparts in some downstream tasks. Further research is necessary to gain a more robust understanding of the effectiveness of different molecular pre-training objectives, so as to provide guidance for optimal methodology design.

Contribution Statement

Jun Xia, Yanqiao Zhu, and Yuanqi Du contribute equally to this work. Stan Z. Li is the corresponding author.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022ZD0115100), the National Natural Science Foundation of China (U21A20427), the Research Center for Industries of the Future (WU2022C043), and the Competitive Research Fund (WU2022A009) from the Westlake Center for Synthetic Biology and Integrated Bioengineering.

References

[Axelrod and Gómez-Bombarelli, 2022] S. Axelrod and R. Gómez-Bombarelli. GEOM, Energy-Annotated Molecular Conformations for Property Prediction and Molecular Generation. *Sci. Data*, 2022.

[Bagal *et al.*, 2021] V. Bagal, R. Aggarwal, et al. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.*, 2021.

[Berman *et al.*, 2000] H. M. Berman, J. Westbrook, et al. The Protein Data Bank. *Nucleic Acids Res.*, 2000.

[Brown *et al.*, 2020] T. B. Brown, B. Mann, et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

[Chithrananda *et al.*, 2020] S. Chithrananda, G. Grand, et al. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv.org*, October 2020.

[Consonni and Todeschini, 2009] V. Consonni and R. Todeschini. *Molecular Descriptors for Chemoinformatics*. 2009.

[Corso *et al.*, 2020] G. Corso, L. Cavalleri, et al. Principal Neighbourhood Aggregation for Graph Nets. In *NeurIPS*, 2020.

[Degen *et al.*, 2008] J. Degen, C. Wegscheid-Gerlach, et al. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, 2008.

[Devlin *et al.*, 2019] J. Devlin, M. Chang, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

[Du *et al.*, 2022a] W. Du, H. Zhang, et al. SE(3) Equivariant Graph Neural Networks with Complete Local Frames. In *ICML*, 2022.

[Du *et al.*, 2022b] Y. Du, T. Fu, et al. MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design. *arXiv.org*, March 2022.

[Edwards *et al.*, 2022] C. Edwards, T. Lai, et al. Translation between Molecules and Natural Language. In *EMNLP*, 2022.

[Fang *et al.*, 2022a] X. Fang, L. Liu, et al. Geometry-Enhanced Molecular Representation Learning for Property Prediction. *Nat. Mach. Intell.*, 2022.

[Fang *et al.*, 2022b] Y. Fang, Q. Zhang, et al. Molecular Contrastive Learning with Chemical Element Knowledge Graph. In *AAAI*, 2022.

[Feng *et al.*, 2022] J. Feng, Z. Wang, et al. MGMAE: Molecular Representation Learning by Reconstructing Heterogeneous Graphs with A High Mask Ratio. In *CIKM*, 2022.

[Gaulton *et al.*, 2012] A. Gaulton, L. J. Bellis, et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.*, 2012.

[Guo *et al.*, 2022] Z. Guo, P. K. Sharma, et al. Multilingual Molecular Representation Learning via Contrastive Pre-training. In *ACL*, 2022.

[Hassani and Ahmadi, 2020] K. Hassani and A. H. K. Ahmadi. Contrastive Multi-View Representation Learning on Graphs. In *ICML*, 2020.

[Hawkins, 2017] P. C. D. Hawkins. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.*, 2017.

[He *et al.*, 2016] K. He, X. Zhang, et al. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[Hjelm *et al.*, 2019] R. D. Hjelm, A. Fedorov, et al. Learning Deep Representations by Mutual Information Estimation and Maximization. In *ICLR*, 2019.

[Ho *et al.*, 2020] J. Ho, A. Jain, et al. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020.

[Honda *et al.*, 2019] S. Honda, S. Shi, et al. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. *arXiv.org*, November 2019.

[Hou *et al.*, 2022] Z. Hou, X. Liu, et al. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *KDD*, 2022.

[Hu *et al.*, 2020a] W. Hu, B. Liu, et al. Strategies for Pre-training Graph Neural Networks. In *ICLR*, 2020.

[Hu *et al.*, 2020b] Z. Hu, Y. Dong, et al. GPT-GNN: Generative Pre-Training of Graph Neural Networks. In *KDD*, 2020.

[Hu *et al.*, 2021] W. Hu, M. Fey, et al. OGB-LSC: A large-scale challenge for machine learning on graphs. In *NeurIPS Datasets and Benchmarks*, 2021.

[Hu *et al.*, 2022] B. Hu, J. Xia, J. Zheng, C. Tan, Y. Huang, Y. Xu, and S. Z. Li. Protein language models and structure prediction: Connection and progression. *arXiv preprint arXiv:2211.16742*, 2022.

[Huang *et al.*, 2021] K. Huang, T. Fu, et al. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. In *NeurIPS Datasets and Benchmarks*, 2021.

[Jiao *et al.*, 2023] R. Jiao, J. Han, et al. Energy-Motivated Equivariant Pretraining for 3D Molecular Graphs. In *AAAI*, 2023.

[Kearnes *et al.*, 2016] S. Kearnes, K. McCloskey, et al. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided Mol. Des.*, 2016.

[Kim *et al.*, 2022] D. Kim, J. Baek, et al. Graph Self-supervised Learning with Accurate Discrepancy Learning. In *NeurIPS*, 2022.

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

[Krenn *et al.*, 2020] M. Krenn, F. Häse, et al. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn. Sci. Technol.*, 2020.

[Li *et al.*, 2021] P. Li, J. Wang, et al. An Effective Self-Supervised Framework for Learning Expressive Molecular Global Representations to Drug Discovery. *Briefings Bioinform.*, 2021.

[Li *et al.*, 2022a] H. Li, D. Zhao, et al. KPGT: Knowledge-Guided Pre-training of Graph Transformer for Molecular Property Prediction. In *KDD*, 2022.

[Li *et al.*, 2022b] S. Li, J. Zhou, et al. GeomGCL: Geometric Graph Contrastive Learning for Molecular Property Prediction. In *AAAI*, 2022.

[Lin *et al.*, 2022] X. Lin, C. Xu, et al. PanGu Drug Model: Learn a Molecule Like a Human. *bioRxiv.org*, April 2022.

[Liu *et al.*, 2022] S. Liu, H. Wang, et al. Pre-training Molecular Graph Representation with 3D Geometry. In *ICLR*, 2022.

[Liu *et al.*, 2023] S. Liu, H. Guo, et al. Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching. In *ICLR*, 2023.

[Nguyen *et al.*, 2021] T. Nguyen, H. Le, et al. GraphDTA: Predicting Drug-Target Binding Affinity with Graph Neural Networks. *Bioinform.*, 2021.

[Pinheiro *et al.*, 2022] G. A. Pinheiro, J. L. Da Silva, et al. SMICLR: Contrastive Learning on Multiple Molecular Representations for Semisupervised and Unsupervised Representation Learning. *J. Chem. Inf. Model.*, 2022.

[Raffel *et al.*, 2020] C. Raffel, N. Shazeer, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 2020.

[Rong *et al.*, 2020] Y. Rong, Y. Bian, et al. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *NeurIPS*, 2020.

[Ross *et al.*, 2022] J. Ross, B. Belgodere, et al. Molformer: Large Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.*, 2022.

[Sanchez-Gonzalez *et al.*, 2020] A. Sanchez-Gonzalez, J. Godwin, et al. Learning to Simulate Complex Physics with Graph Networks. In *ICML*, 2020.

[Satorras *et al.*, 2021] V. G. Satorras, E. Hoogeboom, et al. E(n) Equivariant Graph Neural Networks. In *ICML*, 2021.

[Schütt *et al.*, 2017] K. Schütt, P.-J. Kindermans, et al. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. In *NIPS*, 2017.

[Schütt *et al.*, 2021] K. Schütt, O. T. Unke, et al. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *ICML*, 2021.

[Stärk *et al.*, 2022] H. Stärk, D. Beaini, et al. 3D Infomax Improves GNNs for Molecular Property Prediction. In *ICML*, 2022.

[Sterling and Irwin, 2015] T. Sterling and J. J. Irwin. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.*, 2015.

[Stumpfe *et al.*, 2019] D. Stumpfe, H. Hu, et al. Evolving Concept of Activity Cliffs. *ACS Omega*, 2019.

[Sun *et al.*, 2020] F. Sun, J. Hoffmann, et al. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *ICLR*, 2020.

[Sun *et al.*, 2021] M. Sun, J. Xing, et al. MoCL: Contrastive Learning on Molecular Graphs with Multi-level Domain Knowledge. In *KDD*, 2021.

[Sun *et al.*, 2022] R. Sun, H. Dai, et al. Does GNN Pretraining Help Molecular Representation? In *NeurIPS*, 2022.

[Suresh *et al.*, 2021] S. Suresh, P. Li, et al. Adversarial Graph Augmentation to Improve Graph Contrastive Learning. In *NeurIPS*, 2021.

[Vaswani *et al.*, 2017] A. Vaswani, N. Shazeer, et al. Attention is All You Need. In *NIPS*, 2017.

[Velickovic *et al.*, 2018] P. Velickovic, G. Cucurull, et al. Graph Attention Networks. In *ICLR*, 2018.

[Velickovic *et al.*, 2019] P. Velickovic, W. Fedus, et al. Deep Graph Infomax. In *ICLR*, 2019.

[Wang *et al.*, 2017] Y. Wang, S. H. Bryant, et al. Pubchem Bioassay: 2017 Update. *Nucleic Acids Res.*, 2017.

[Wang *et al.*, 2019] S. Wang, Y. Guo, et al. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *BCB*, 2019.

[Wang *et al.*, 2021] Y. Wang, Y. Min, et al. Molecular Graph Contrastive Learning with Parameterized Explainable Augmentations. In *BIBM*, 2021.

[Wang *et al.*, 2022a] Y. Wang, R. Magar, et al. Improving Molecular Contrastive Learning via Faulty Negative Mitigation and Decomposed Fragment Contrast. *J. Chem. Inf. Model.*, 2022.

[Wang *et al.*, 2022b] Y. Wang, J. Wang, et al. MolCLR: Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.*, 2022.

[Weininger, 1988] D. Weininger. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, 1988.

[Wu *et al.*, 2018] Z. Wu, B. Ramsundar, et al. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.*, 2018.

[Xia *et al.*, 2022a] J. Xia, L. Wu, et al. ProGCL: Rethinking Hard Negative Mining in Graph Contrastive Learning. In *ICML*, 2022.

[Xia *et al.*, 2022b] J. Xia, L. Wu, et al. SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Augmentation. In *WWW*, 2022.

[Xia *et al.*, 2023] J. Xia, C. Zhao, et al. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. In *ICLR*, 2023.

[Xu *et al.*, 2019] K. Xu, W. Hu, et al. How Powerful are Graph Neural Networks? In *ICLR*, 2019.

[Xu *et al.*, 2021a] D. Xu, W. Cheng, et al. InfoGCL: Information-Aware Graph Contrastive Learning. In *NeurIPS*, 2021.

[Xu *et al.*, 2021b] M. Xu, H. Wang, et al. Self-supervised Graph-level Representation Learning with Local and Global Structure. In *ICML*, 2021.

[Xu *et al.*, 2021c] Z. Xu, Y. Luo, et al. Molecule3D: A Benchmark for Predicting 3D Geometries from Molecular Graphs. *arXiv.org*, September 2021.

[Yi *et al.*, 2022] J. Yi, C. Wu, et al. MICER: A Pre-Trained Encoder–Decoder Architecture for Molecular Image Captioning. *Bioinform.*, 2022.

[Ying *et al.*, 2021] C. Ying, T. Cai, et al. Do Transformers Really Perform Badly for Graph Representation? In *NeurIPS*, 2021.

[You *et al.*, 2020] Y. You, T. Chen, et al. Graph Contrastive Learning with Augmentations. In *NeurIPS*, 2020.

[You *et al.*, 2021] Y. You, T. Chen, et al. Graph Contrastive Learning Automated. In *ICML*, 2021.

[You *et al.*, 2022] Y. You, T. Chen, et al. Bringing Your Own View: Graph Contrastive Learning without Prefabricated Data Augmentations. In *WSDM*, 2022.

[Zaidi *et al.*, 2023] S. Zaidi, M. Schaarschmidt, et al. Pre-training via Denoising for Molecular Property Prediction. In *ICLR*, 2023.

[Zeng *et al.*, 2022a] X. Zeng, H. Xiang, et al. Accurate Prediction of Molecular Properties and Drug Targets Using a Self-Supervised Image Representation Learning Framework. *Nat. Mach. Intell.*, 2022.

[Zeng *et al.*, 2022b] Z. Zeng, Y. Yao, et al. A Deep-Learning System Bridging Molecule Structure and Biomedical Text with Comprehension Comparable to Human Professionals. *Nat. Commun.*, 2022.

[Zhang *et al.*, 2021] Z. Zhang, Q. Liu, et al. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. In *NeurIPS*, 2021.

[Zheng *et al.*, 2022] J. Zheng, Y. Wang, G. Wang, J. Xia, Y. Huang, G. Zhao, Y. Zhang, and S. Z. Li. Using context-to-vector with graph retrofitting to improve word embeddings. *ACL*, 2022.

[Zhou *et al.*, 2023] G. Zhou, Z. Gao, et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *ICLR*, 2023.

[Zhu *et al.*, 2021] J. Zhu, Y. Xia, et al. Dual-view Molecule Pre-training. *arXiv.org*, June 2021.

[Zhu *et al.*, 2022a] J. Zhu, Y. Xia, et al. Unified 2D and 3D Pre-Training of Molecular Representations. In *KDD*, 2022.

[Zhu *et al.*, 2022b] Y. Zhu, D. Chen, et al. Featurizations Matter: A Multiview Contrastive Learning Approach to Molecular Pre-training. In *AI4Science@ICML*, 2022.