SCIENCE CHINA
Information Sciences

• REVIEW •

July 2025, Vol. 68, Iss. 7, 170104:1–170104:19
https://doi.org/10.1007/s11432-024-4457-2

Special Topic: AI for Biology

Molecular pretraining models towards molecular property prediction

Jianbo QIAO[1], Wenjia GAO[1], Junru JIN[1], Ding WANG[1], Xu GUO[1], Balachandran MANAVALAN[2*] & Leyi WEI[3,4*]

[1]*School of Software, Shandong University, Jinan 250101, China*
[2]*Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon 16419, Republic of Korea*
[3]*Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China*
[4]*Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan 250101, China*

Received 31 December 2024/Revised 28 February 2025/Accepted 9 March 2025/Published online 23 June 2025

Abstract Molecular property prediction plays a pivotal role in advancing our understanding of molecular representations, serving as a key driver for progress in drug discovery. Leveraging deep learning to gain comprehensive insights into molecular properties has become increasingly critical. Recent breakthroughs in molecular property prediction have been achieved through molecular pretraining models, which utilize large-scale databases of unlabeled molecules for pretraining, followed by fine-tuning for specific downstream tasks. These models enable a deeper understanding of molecular properties. In this study, we review recent advancements in molecular property prediction using molecular pretraining models. Our focus includes molecular descriptors, the impact of pretraining dataset size, molecular characterization model architectures, and the diversity of pretraining task types. Additionally, we compare the performance of existing methods and propose future directions to enhance the effectiveness of molecular pretraining models.

Keywords molecular pretraining models, molecular property prediction, graph neural network (GNN), graph Transformer, PubChem, MoleculeNet

Citation Qiao J B, Gao W J, Jin J R, et al. Molecular pretraining models towards molecular property prediction. Sci China Inf Sci, 2025, 68(7): 170104, https://doi.org/10.1007/s11432-024-4457-2

1 Introduction

Deep learning, with its powerful learning capabilities, can process raw molecular representations and automatically extract task-relevant features from vast datasets, enabling highly accurate predictions [1–6]. This makes deep learning models particularly well-suited for large-scale drug screening, where their efficient runtime significantly accelerates the screening process and offers novel insights into disease mechanisms [7–11]. Despite these advantages, the complexity of molecular characterization poses challenges, especially when training high-dimensional molecular feature representations from scratch with limited data [12–15]. To overcome this, molecular pretraining models (MPMs) have been introduced. These models leverage large-scale unlabeled molecular datasets to learn generalizable molecular representations, enabling more accurate molecular property predictions.

Chemical molecules can be represented in various forms, including sequence-based simplified molecular input line entry system (SMILES) strings [16,17], 2D molecular diagrams or images [18,19], and 3D molecular conformations or videos [20]. To leverage these representations, a range of pretraining tasks has been developed, such as atom masking, interatomic distance prediction, and functional group prediction. Additional tasks include node-level atomic attribute prediction, graph-level molecular motif prediction, molecular contrastive learning, and 3D coordinate recovery that incorporates spatial structural information. These pretraining strategies are diverse and intricate, necessitating explicit analysis to refine and enhance their effectiveness.

As research advances, molecular pretraining models increasingly integrate 3D conformational information to improve predictive performance. The evolution of molecular encoders has progressed from

* Corresponding author (email: bala2022@skku.edu, weileyi@sdu.edu.cn)

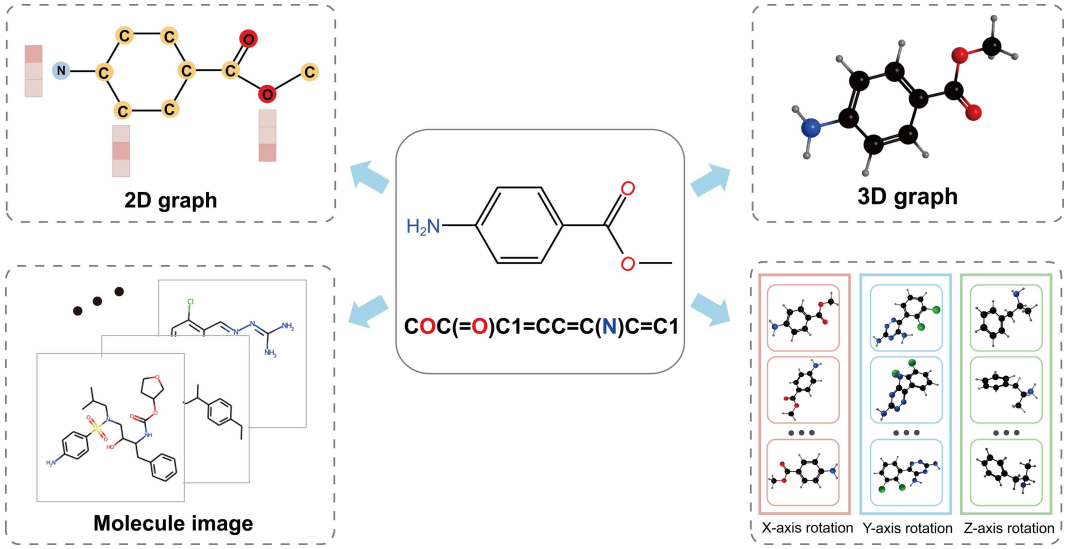

Figure 1 (Color online) Different types of molecular descriptors.

sequence-based models, such as recurrent neural networks (RNNs) and Transformers, to graph neural networks (GNNs) [21,22]. Notably, recent studies highlight the superior performance of graph Transformer-based molecular encoders in molecular property prediction tasks. Their ability to capture global features and fuse multimodal information has demonstrated significant advantages in advancing molecular representation learning.

In this study, we provide a comprehensive overview of the development of molecular pretraining models, focusing on molecular descriptors, pretraining dataset sizes, molecular representation model architectures, and types of pretraining tasks. This review offers an organized perspective on the topic. We also compare the performance of various methods in molecular property prediction tasks using both scaffold split [23] and random scaffold split [24] evaluations, offering valuable insights for method assessment. Finally, we identify key research challenges in molecular pretraining models and propose future directions to advance this field further.

2 Molecular descriptors

Chemical molecules can be described using various molecular descriptors, including sequence-based representations, 2D representations such as molecular diagrams and images, and 3D representations like molecular conformations and videos (Figure 1). This section provides an in-depth discussion of these diverse molecular descriptors.

2.1 Sequence-based molecular representation

Chemical molecules can be represented as strings using the SMILES. In SMILES strings, atoms and chemical bonds are denoted by letters and punctuation marks, respectively, with branches indicated by parentheses. This form of molecular representation is well-suited for natural language processing techniques due to its sequential structure. While SMILES is compact and computationally efficient, it has notable limitations.

First, the same molecule can have multiple valid SMILES representations, undermining the uniqueness of molecular representations and complicating molecular representation learning. Second, while SMILES can capture a molecule's 2D topology to some extent, it cannot represent the 3D spatial structure or account for the conformational diversity of multi-conformational molecules.

SELF-referencing embedded strings (SELFIES [25]) is an improved version of SMILES that ensures all generated strings correspond to valid chemical structures, thereby preventing training failures caused by invalid SMILES. However, SELFIES has low readability. In contrast, the International Union of Pure and Applied Chemistry (IUPAC [26]) nomenclature system provides a systematic approach to describing molecular structures and functional groups in a human-readable format.

Alternatively, molecular fingerprints provide another sequence-based representation for molecules, encoding structural information into fixed-length vectors [27–30]. These vectors algorithmically summarize chemical information, such as atoms, bonds, and characteristic substructures. For instance, MACCS [31] fingerprints encode 2D substructures, while extended connectivity fingerprints (ECFPs [32]) are based on topological features, such as ring neighborhoods. These fingerprints are widely used in drug design and quantitative structure-activity relationship (QSAR) analysis due to their ability to efficiently represent molecular features.

2.2 2D-based molecular representation

Graph data naturally model complex interactions and are extensively used in natural sciences, including chemistry and biology, as well as in fields such as image processing and social network analysis. Graph-based molecular representations, where atoms are treated as nodes and chemical bonds as edges, enable chemical molecules to be represented as 2D graph structures. These 2D topological graphs effectively capture the topological features of molecular structures, providing an intuitive understanding of molecular organization. However, they are limited in their ability to represent 3D conformational information [33].

Recently, molecular representation through molecular images has emerged as a novel and effective approach [18]. In this method, chemical molecules are visualized as 2D images or pixel-based representations in two-dimensional space. Leveraging computer vision techniques, this approach facilitates molecular representation learning tasks, such as calculating intermolecular similarity and extracting key substructures, broadening the scope of molecular characterization.

2.3 3D-based molecular representation

Molecular graphs are effective for modeling the 2D features of molecules but are limited in their ability to capture 3D spatial information, such as molecular conformations. Additionally, the complexity and diversity of molecular graphs—such as heterogeneity and long-range interactions in varying chemical environments—pose challenges for 2D-based molecular representation learning [34]. To address these limitations, 3D spatial representations of molecules have been introduced for molecular representation learning.

The 3D information of molecules includes atomic spatial coordinates, interatomic distances, bond angles, and other conformational details. By incorporating 3D information, optimal molecular conformations can be selected, making this approach well-suited for high-precision chemical modeling and complex molecular simulations. Moreover, 3D data enable the calculation of quantum chemical (QC) properties, such as energy, polarizability, and molecular orbitals, enhancing the interpretability and predictive power of molecular representations. However, 3D structural information comes with challenges, including high computational complexity, costly data generation, and sensitivity to conformational energy variations that significantly impact molecular characterization.

Beyond static 3D representations, molecular spatial structures can also be visualized as dynamic videos [20]. By rendering a molecule's 3D structure as a series of frames, this novel approach captures temporal and spatial conformational information, offering new opportunities for molecular representation learning.

3 Datasets and split methods

3.1 Pretraining datasets

PubChem [35], an open chemistry database from the National Institutes of Health (NIH), contains over 150 million compounds and provides diverse molecular information, including chemical structures, identifiers, chemical and physical properties, biological activities, patents, and safety and toxicity data.

The ZINC [36] database hosts over 200 million molecules organized into subsets such as drug-like, lead-like, and natural products. It offers 3D conformations suitable for molecular docking and molecular dynamics simulations. ZINC15, an enhanced version of ZINC, features a larger dataset, improved flexibility, and optimized molecular information tailored for modern virtual screening applications. ZINC20 expands upon its predecessor by incorporating over 1.7 billion molecules, available in SMILES, SDF, Mol2, PDBQT, and other formats.

Table 1 Pretraining datasets employed by various molecular pretraining methods, along with the number of molecules used for pretraining. k for thousands and M for millions.

Method	Datasets	Number of molecules
KPGT	ChEMBL	2M
FG-BERT	ChEMBL	1.45M
PremuNet	ChEMBL, PCQM4Mv2	5M
MolGT	GEOM-Drugs, QMugs, PCQM4Mv2	6M
GeoSSL-DDM	Molecule3D [41]	1M
Transformer-M	PCQM4Mv2	3.37M
Uni-Mol+	PCQM4Mv2	3.37M
VideoMol	PCQM4Mv2	2M
GNS-TAT	PCQM4Mv2	3.37M
Frad	PCQM4Mv2	3.37M
MolCLR	PubChem	10M
ChemBERTa	PubChem	77M
ChemBERTa-2	PubChem	77M
ImageMol	PubChem	10M
CAFE-MPP	PubChem	10M
BioT5+	PubChem, ZINC20	28.8M
Chemformer	ZINC15	100M
SMILES-BERT	ZINC15	18.6M
GEM	ZINC15	20M
KANO	ZINC15	0.25M
MGSSL	ZINC15	250k
HiMol	ZINC15	250k
Uni-Mol	ZINC15, ChEMBL	19M
Mol-AE	ZINC15, ChEMBL	19M
BioT5	ZINC20	–

ChEMBL [37], focused on drug discovery, includes more than 2.5 million bioactive chemical structures. This high-quality resource links bioactive compounds with drug targets, supporting both academic and industrial research in identifying and developing new drugs.

The PCQM4Mv2 [38], part of the PubChemQC project, is a quantum chemistry dataset comprising 3378606 molecules with 3D structures calculated via density functional theory (DFT).

GEOM-Drugs [39] provides 3D conformations of a vast number of molecules generated through quantum chemical optimization or force-field-based methods. Its molecular library is primarily derived from ZINC and ChEMBL, ensuring drug-like properties that make it well-suited for drug discovery and virtual screening studies. Similarly, QMugs [40] selects drug-like compounds from the PubChem database and computes high-precision quantum chemical properties for these molecules.

The pretraining datasets used for the methods mentioned in the text are given in Table 1 [41].

3.2 Downstream task datasets

MoleculeNet [42] is a benchmark dataset specifically designed for molecular machine learning tasks. It includes classification tasks, such as BACE, BBBP, ClinTox, Tox21, HIV, SIDER, MUV, and ToxCast, as well as regression tasks like FreeSolv, ESOL, and Lipophilicity. MoleculeNet encompasses a wide range of challenges, from target prediction to drug toxicity and solubility assessment (Table 2).

Two widely used dataset splitting strategies in MoleculeNet are scaffold split [23] and random scaffold split [24]: scaffold split based on molecular scaffolds, ensuring that scaffolds in the training, validation, and test sets are distinct. By simulating out-of-distribution prediction tasks, scaffold split better reflects real-world scenarios where models must generalize to novel chemical structures. Random scaffold split combines elements of random and scaffold-based splitting, this method introduces a degree of randomness to the assignment of molecules while retaining the scaffold-based separation logic. Unlike standard scaffold splitting, where scaffold groupings are fully assigned to a single dataset (training, validation, or test), random scaffold split proportionally distributes scaffolds across the datasets.

The scaffold split is more challenging for model evaluation as it creates a stricter test of a model's ability to generalize to unseen scaffolds, while the random scaffold split offers greater flexibility by blending

Table 2 Details of the molecular property datasets.

Dataset	Number of molecules	Number of tasks	Category	Metric
BACE	1513	1	Biophysics	ROC-AUC
BBBP	2039	1	Physiology	ROC-AUC
ClinTox	1478	2		
Tox21	7831	12		
ToxCast	8575	617		
SIDER	1427	27		
MUV	93087	17	Biophysics	ROC-AUC
MIV	41127	1		
FreeSolv	642	1	Physical chemistry	RMSE
ESOL	1128	1		
Lipophilicity	4200	1		
QM7	6830	1	Quantum mechanics	MAE
QM8	21786	12		
QM9	133885	3		

scaffold logic with random assignment.

According to MoleculeNet [42], the area under the receiver operating characteristic curve (ROC-AUC) [43] is employed to evaluate the performance of classification tasks, while the root mean square error (RMSE) [44] or the mean absolute error (MAE) is employed to evaluate the performance of regression tasks.

BACE [45] dataset provides quantitative (IC50) and qualitative (binary label) binding results for a set of inhibitors of human β-secretase 1 (BACE-1).

BBBP [46] dataset is extracted from a study on the modeling and prediction of the barrier permeability.

ClinTox [47] dataset compares drugs approved by the FDA and drugs that have failed clinical trials for toxicity reasons.

Tox21 [48] dataset contains qualitative toxicity measurements for 8k compounds on 12 different targets, including nuclear receptors and stress response pathways.

ToxCast [49] dataset is an extended data collection from the same initiative as Tox21, providing toxicology data for a large library of compounds based on in vitro high-throughput screening. The processed collection includes qualitative results of over 600 experiments on 8k compounds.

SIDER [50] is a database of marketed drugs and adverse drug reactions (ADR), grouped into 27 system organ classes.

MUV [51] group is a benchmark dataset selected from PubChem BioAssay by applying a refined nearest neighbor analysis. The MUV dataset contains 17 challenging tasks of approximately 90000 compounds, designed specifically to validate virtual screening techniques.

HIV [52] dataset contains more than 40000 records of whether the compound inhibits HIV replication for binary classification between active and inactive.

FreeSolv [53] provides experimental and calculated hydration free energy of small molecules in water. The calculated values are derived from alchemical free energy calculations using molecular dynamics simulations. The experimental values are included in the benchmark collection.

ESOL [54] is a standard regression dataset containing structures and water solubility data for 1128 compounds. The dataset is widely used to validate machine learning models on estimating solubility directly from molecular structures (as encoded in SMILES strings).

Lipophilicity [55] is a dataset curated from ChEMBL database containing experimental results on octanol/water distribution coefficient (logD at pH = 7.4). Due to the importance of Lipophilicity in membrane permeability and solubility, the task is of high importance to drug development.

QM7 [56] is a subset of GDB-13 (a database of nearly 1 billion stable and synthesizable organic molecules) that records the calculated atomization energies of stable and synthesizable organic molecules, such as HOMO/LUMO and atomization energies. It contains various molecular structures (such as triple bonds, cycles, amides and epoxy resins) and up to 7 heavy atoms C, N, O, and S.

QM8 [57] uses a variety of quantum mechanics methods to calculate the electronic spectrum and excited state energy of small molecules.

QM9 [58] is a comprehensive dataset providing geometric, energetic, electronic, and thermodynamic

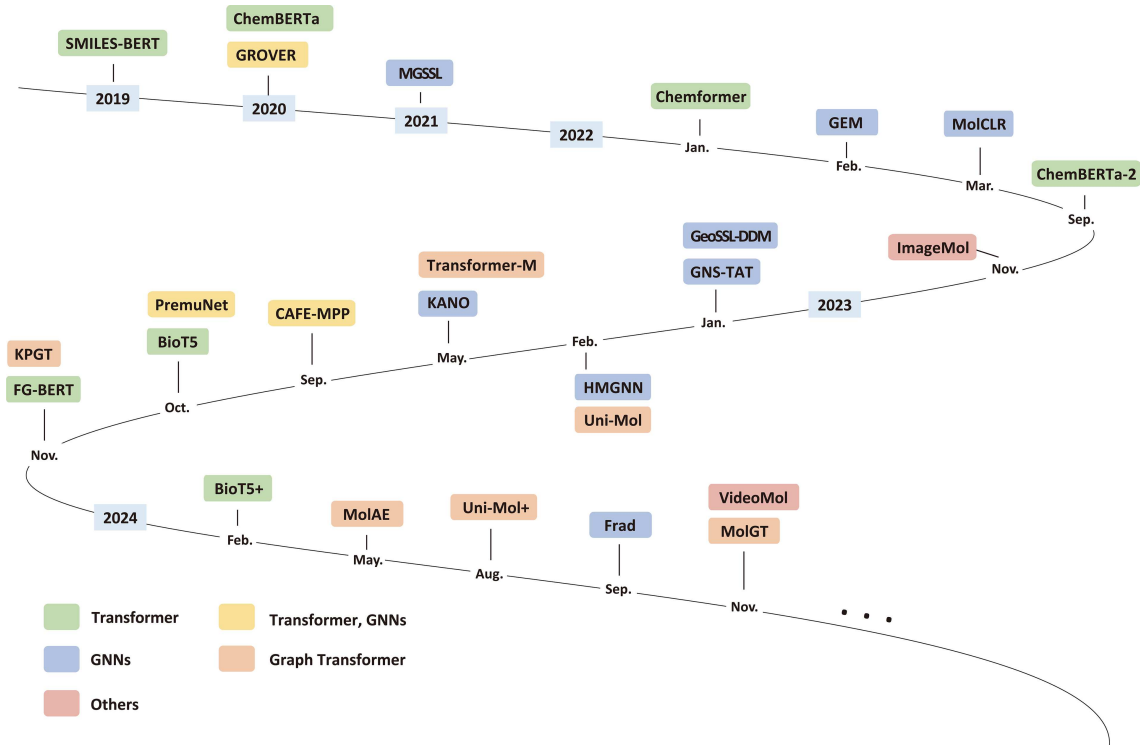

Figure 2 (Color online) Summary of representative molecular pre-training models (MPMs).

properties for a subset of the GDB-17 database, including 134000 stable organic molecules and up to 9 heavy atoms.

4 Molecular encoder architecture

With the development of chemoinformatics, researchers have proposed more and more models for molecular characterization. For the three molecular descriptors mentioned above, we summarize the corresponding model architectures: (1) sequences-based; (2) GNN-based; (3) graph Transformer-based; (4) others. The methods mentioned in this review are summarized in Figure 2 and Table 3 [18, 20, 23, 34, 59–80].

4.1 Sequence-based approaches

Sequence-based approaches draw inspiration from natural language processing (NLP) for molecular data analysis. These methods use sequence-type molecular descriptors, such as SMILES strings and molecular fingerprints, as input to train models based on architectures like RNNs and Transformers for molecular representation learning (Figure 3).

Wang et al. [59] developed SMILES-BERT, a semi-supervised model featuring a Transformer layer built on an attention mechanism. The model was pre-trained on large-scale unlabeled data using the Masked SMILES Recovery task.

Chithrananda et al. [60] and Ahmad et al. [62] developed ChemBERTa and ChemBERTa-2, respectively, leveraging the Transformer architecture for SMILES-based molecular characterization. These models were pre-trained on a dataset of 77 million SMILES molecules, marking one of the first successful applications of Transformer models in this domain.

Chemformer, introduced by Irwin et al. [61], leverages the BART [81] language model by integrating both encoder and decoder architectures of the Transformer and improves molecular representations through masked pretraining.

Zhang et al. [74] integrated molecular fingerprinting with SMILES and innovatively mapped each atom in a SMILES string to its corresponding position in a 2D molecular representation.

FG-BERT, proposed by Li et al. [64], incorporates functional group information and leverages the

Table 3 Details of the model architectures.

Model	Architecture	Date	Molecular descriptors	Datasets	Pretraining strategies
SMILES-BERT [59]	Transformer	2019.09	SMILES	ZINC15	Masked SMILES prediction
ChemBERTa [60]	Transformer	2020.10	SMILES, SELFIES	PubChem	Masked prediction
Chemformer [61]	Transformer	2022.01	SMILES	ZINC15	(1) Masking (2) Augmentation a combination of masking and augmentation
ChemBERTa-2 [62]	Transformer	2022.09	SMILES	PubChem	(1) Masked language modeling (2) Properties prediction
BioT5 [63]	Transformer	2023.10	SELFIES	PubChem, ZINC20	Masked language model
FG-BERT [64]	Transformer	2023.11	SMILES, 2D graph	ChEMBL	(1) Masked language model (2) Next sentence prediction task
BioT5+ [65]	Transformer	2024.02	SELFIES, IUPAC	ZINC20	Masked language model
MGSSL [23]	GNNs	2021.10	2D graph	ZINC15	(1) Masked pretraining for atom attributes (2) Construction and prediction of motif tree
GEM [34]	GNNs	2022.02	3D graph	ZINC15	(1) Predicting bond lengths (2) Predicting bond angles (3) Predicting interatomic distances
MolCLR [66]	GNNs	2022.03	2D graph	PubChem	(1) Molecule graph augmentation (2) Contrastive-based pre-training
GNS-TAT [67]	GNNs	2023.01	3D graph	PCQM4Mv2	Atom 3D coordinate recovery
GeoSSL-DDM [68]	GNNs	2023.01	3D graph	Molecule3D	(1) Atomic 3D coordinate denoising (2) Interatomic distance denoising
HiMol [69]	GNNs	2023.02	2D graph	ZINC15	Node or edge level: (1) Atom type (2) Bond link predection (3) Bond type (single/double bond) Graph level: (4) Atom number (5) Bond number
KANO [70]	GNNs	2023.05	2D graph	ZINC15	Contrastive-based pre-training
Frad [71]	GNNs	2024.09	3D graph	PCQM4Mv2	Coordinate Gaussian noise recover
GROVER [72]	Transformer, GNNs	2020.10	2D graph	ZINC15, ChEMBL	(1) Contextual property prediction (2) Graph-level motif prediction
CAFE-MPP [73]	Transformer, GNNs	2023.09	SMILES, 2D graph	PubChem	Comparative learning based on SMILES and 2D graphs
PremuNet [74]	Transformer, GNNs	2023.10	SMILES, 2D graph, 3D graph	ChEMBL, PCQM4Mv2	(1) AutoEncoder (2) Masked prediction (3) Reconstruction of atomic information based on 3D structures (4) Reconstruction of 3D structures based on atomic information (5) Random masking of 2D and 3D information and reconstruction
Uni-Mol [75]	Graph Transforme	2023.02	3D graph, Protein Pockets	ZINC, ChEMBL	(1) Atom 3D coordinate recovery (2) Atom-atom pair distance prediction (3) Masked pretraining for atom species
Transformer-M [76]	Graph Transforme	2023.05	2D graph, 3D graph	PCQM4Mv2	(1) Prediction HOMO-LUMO gap (2) 3D position denoising
KPGT [77]	Graph Transforme	2023.11	2D graph	PCQM4Mv2	Masked pretraining
Mol-AE [78]	Graph Transforme	2024.05	3D graph	ZINC, ChEMBL	3D cloze
Uni-Mol+ [79]	Graph Transforme	2024.08	3D graph	PCQM4Mv2	(1) QC property prediction (2) 3D position prediction
MolGT [80]	Graph Transforme	2024.11	2D graph, 3D graph	GEOM-Drugs, QMugs, PCQM4Mv2	(1) InfoMotif (2) Coordinate denoising (3) Knowledge-guided prototypical clustering (4) Implicit 3D geometry contrastive learning
ImageMol [18]	Others	2022.11	2D image	PubChem	(1) Molecular image reconstruction (2) Mask-based contrastive learning (3) Molecular rationality discrimination (4) Jigsaw puzzle prediction
VideoMol [20]	Others	2024.11	3D video	PCQM4Mv2	(1) Video-aware pre-training (2) Direction-aware pre-training (3) Chemical-aware pre-training

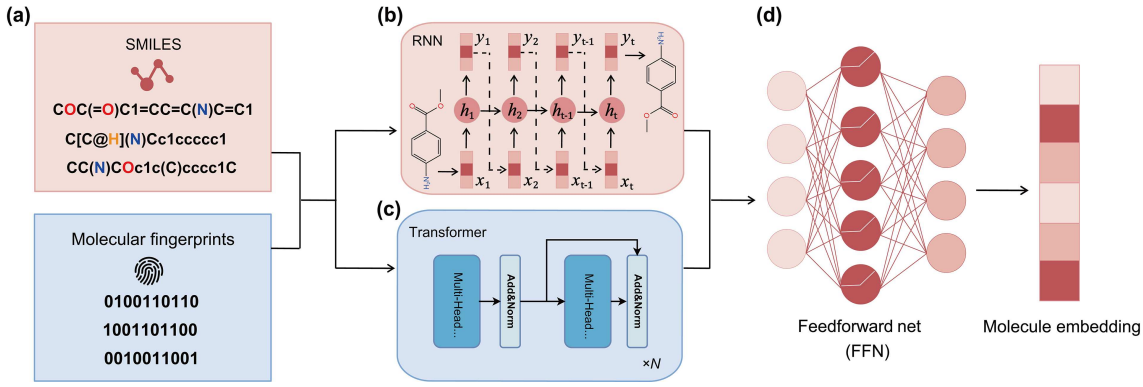

Figure 3 (Color online) Processing of sequence-based molecular representation using RNN and Transformer architecture. (a) Representation of molecules using SMILES or molecular fingerprints; (b) RNN architecture; (c) Transformer architecture; (d) obtaining molecular embedding.

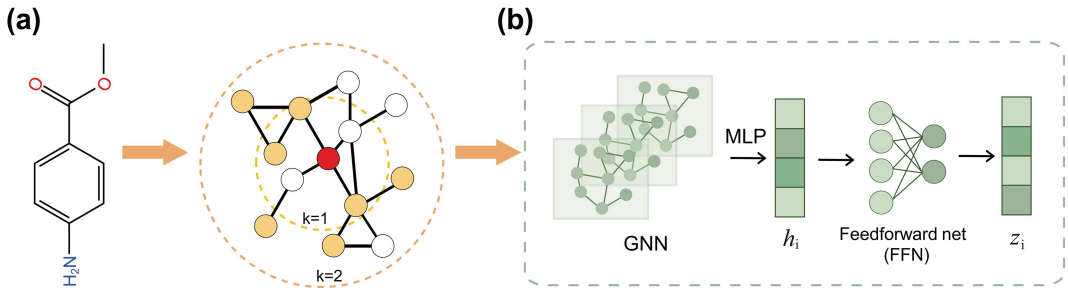

Figure 4 (Color online) Processing of molecular representation using GNN architecture. (a) Convert chemical molecules into molecular graphs and update node representations through a message-passing mechanism, where k denotes k-hop neighbors; (b) after passing through a multilayer GNN, the graph-level molecular representation h_i is obtained, followed by the final representation z_i produced through a feed-forward neural network.

attention mechanism to emphasize FG features critical to target attributes, thereby offering strong interpretability for downstream training tasks.

The 1D SMILES sequence-based molecular characterization method obtains a low-dimensional vector representation of the molecule by performing feature extraction on the SMILES string. However, the performance of many SMILES-based deep neural network models is constrained by the limitations of poor scalability, loss of spatial information, and non-uniqueness of the SMILES-ordered representation. Therefore, SELFIES and IUPAC nomenclature were introduced for molecular representation.

With the growing research on large language models (LLMs), multimodal pre-trained models integrating diverse information have been applied to molecular property prediction. Beyond property prediction, these studies on the sequence-based representation of molecules offer new insights into molecular characterization. Pei et al. [63] introduced BioT5, which leverages the SELFIES representation to extract molecular features. They employed the T5 (text-to-text transfer Transformer) architecture with an encoder-decoder structure to process SELFIES. Later, they proposed BioT5+ [65], which further incorporated IUPAC nomenclature to enhance molecular characterization by jointly utilizing SELFIES and IUPAC.

4.2 GNN-based approaches

GNN-based modeling architectures are commonly employed to process 2D molecular graphs or 2D-like molecular graphs augmented with 3D information. These models utilize message passing within the graph to learn the properties of nodes (atoms) and edges (bonds), effectively capturing both local and global structural information (Figure 4). GNNs are particularly well-suited for tasks involving complex molecular structures and strong interdependencies, such as molecular property prediction and reaction prediction.

Classical graph neural network models, such as GCN [82], GAT [83], GIN [84], MPNN [85], and GraphSAGE [86], can model molecular structures to varying extents. Building on these, Song et al. [87] proposed CMPNN, which introduces a new molecular embedding method to enhance message interaction

between nodes and edges. This is achieved through the incorporation of a message enhancer module, addressing challenges in the representation of molecular graphs. Although the models mentioned above are trained from scratch without a pretraining process, they offer valuable insights [64] for the development of pre-trained models.

Wang et al. [66] introduced MolCLR, which employs GNN encoders based on GCN and GIN. By utilizing both original molecular graphs and masked augmented graphs in a comparative learning strategy, MolCLR improves molecular representation learning. However, these models remain constrained to the 2D topology of molecules and neglect 3D spatial information.

To overcome this limitation, Fang et al. [34] proposed GEM, a molecular representation model that integrates both 2D and 3D molecular information. They restructured the molecular graph into a bond angle diagram, where bond angles are treated as edges and bonds as atoms. This enables a "3D modeling" approach through pretraining tasks designed to incorporate spatial structure.

GNS-TAT, proposed by Zaidi et al. [67], exclusively utilizes 3D molecular graphs to learn specific force fields through 3D structural space, relying on the atomic 3D coordinates for knowledge acquisition. Building on this, Liu et al. [68] were inspired by the dynamic properties of 3D molecules, where molecular motion in 3D Euclidean space creates a smooth potential surface. They introduced 3D interatomic distances and proposed GeoSSL-DDM. Similarly, Ni et al. [71] developed the Frad framework, which incorporates noise design and refines the processing of 3D molecular graphs using GNNs, enhancing the representation of 3D molecular structures.

Additionally, Zhang et al. [23] proposed MGSSL, which converts molecular graphs into motif trees based on key substructures within molecules and employs GNNs to capture graph-level knowledge at the motif level. Similarly, Zang et al. [69] introduced HiMol, which constructs a three-level graph representation encompassing nodes, motifs, and the overall graph, enabling comprehensive integration of motif information with molecular graph data. However, their approach defines motifs in a broad sense, lacking specificity to functional groups in the strict sense. Fang et al. [70] introduced KANO, a model based on knowledge graphs and functional group prompts. By constructing a knowledge graph based on chemical elements and enhancing the molecular graph with knowledge graph embeddings and functional group prompts, KANO aids in identifying key substructures within molecules, enhancing understanding of molecular properties. CAFE-MPP, proposed by Xie et al. [73], facilitates multi-view interactions through contrastive learning between two modalities: SMILES and molecular graph. Later, Zhang et al. [74] proposed PremuNet, which integrated feature fusion across three molecular modalities-1D, 2D, and 3D-further advancing the development of GNNs.

While GNNs are adept at updating node and edge features through message passing, they face limitations when applied to complex graphs. Specifically, the limited range of message passing can hinder the capture of global information between distant nodes. Moreover, GNNs may suffer from the oversmoothing problem, where repeated updates result in the homogenization of node features, diminishing the model's ability to differentiate between distinct molecular structures.

4.3 Graph Transformer-based approaches

Graph-based Transformer molecular representations leverage the strengths of the Transformer architecture, originally designed for sequence data, to process molecular graphs with complex dependencies. Unlike traditional GNNs, graph Transformers utilize a self-attention mechanism to compute dependencies between nodes, enabling them to effectively capture long-range structural information. This makes them particularly suited for large-scale molecular datasets, offering greater flexibility and expressiveness in handling diverse molecular structures (Figure 5).

One notable model, GROVER [72], combines dynamic message passing networks with Transformer-like architecture but retains GNN components, making it a hybrid rather than a pure Transformer approach.

In contrast, Graphormer fully abandons the message-passing paradigm, introducing innovations such as distance encoding based on shortest paths between nodes, edge encoding to include bond-specific information, and node centrality encoding to represent the importance of nodes within the graph. Additionally, the use of a virtual node allows for global information flow across the entire graph, enabling a comprehensive understanding of 2D molecular structures.

Building on Graphormer, Transformer-M [76] introduced a dual-channel model designed for both 2D and 3D molecular tasks. For 2D molecular graphs, it incorporates shortest paths, edge attributes, and

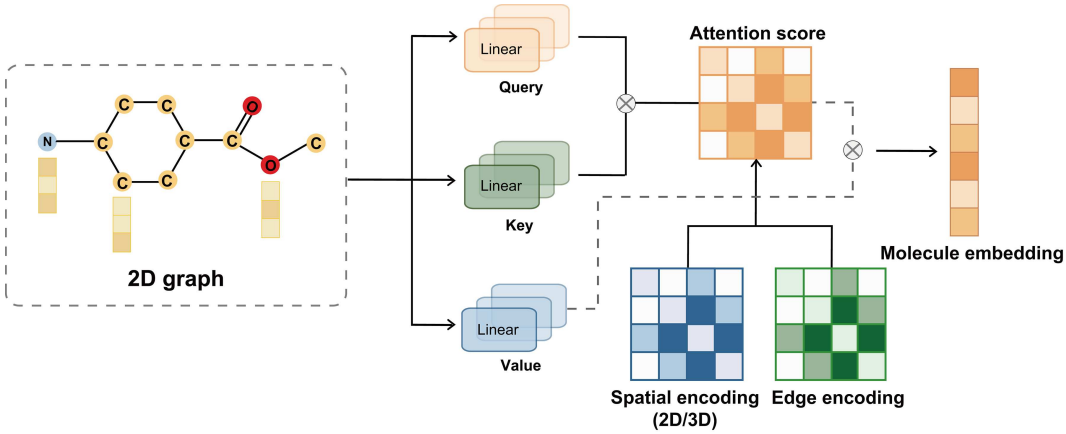

Figure 5 (Color online) Processing of molecular representation using graph Transformer architecture. The self-attention mechanism enables the graph Transformer to capture global dependencies, while incorporating spatial and edge encodings as biases to the attention scores enhances the model's ability to capture both local and global relationships within molecular graphs.

node degree encoding, while for 3D molecular graphs, it integrates spatial distances derived from atomic coordinates, using these as attention biases to guide the learning of inter-atomic dependencies.

Expanding on this foundation, Chen et al. [80] developed the modal-sharing graph Transformer to enhance knowledge sharing between 2D and 3D molecular features. Uni-Mol [75] adds atom-to-atom 3D distance coding with noise-enhanced attention biases to improve the model's grasp of 3D information. Uni-Mol+ [79] further advances this by employing mechanisms like outer product and triangular update to facilitate interactions between molecular encodings and the 3D distance matrix, dynamically updating this matrix at each Transformer layer to enhance structural understanding. Mol-AE [78], based on Uni-Mol, takes a different approach by using a self-encoder to reconstruct molecular latent representations into 3D molecular information, minimizing reconstruction loss.

Li et al. [77] introduced the line graph Transformer (LiGhT), which captures complex patterns in molecular graph structures without relying on 3D spatial information. Their approach extends each molecular graph by adding a knowledge node connected to the original nodes, allowing the backbone model to effectively capture both structural and semantic information within the molecular graph.

Despite their capabilities, graph Transformers face challenges due to their high computational complexity, which limits scalability and makes processing large molecules or highly intricate structures difficult. Nonetheless, their ability to flexibly adapt to the multi-scale features of complex graphs and incorporate 3D conformational information during 2D molecular graph processing represents a significant advancement in molecular representation learning. Balancing their computational demands with scalability remains an ongoing area of research, aiming to optimize these powerful models for broader applications.

4.4 Other approaches

Molecular images can be generated from chemical structures using tools like RDKit, capturing the 2D molecular structure and enabling the incorporation of functional group information through image recognition techniques. This approach has led to innovative frameworks for molecular representation using computer vision. Zeng et al. [18] introduced ImageMol, an unsupervised molecular image pre-training framework that incorporates chemical-awareness capabilities. By representing compounds as molecular images and employing ResNet as a molecular encoder, ImageMol effectively extracted latent molecular features. Building on this, Xiang et al. [20] proposed VideoMol, which leveraged PyMOL to render 60-frame motion videos for each 3D molecule, capturing dynamic structural information. Using vision Transformers (ViT) to extract molecular representations, VideoMol extended the use of computer vision from static 2D images to dynamic 3D molecular videos. Together, these methods highlight the potential of utilizing 2D molecular images and 3D molecular videos as novel descriptors to enhance molecular representation learning through computer vision.

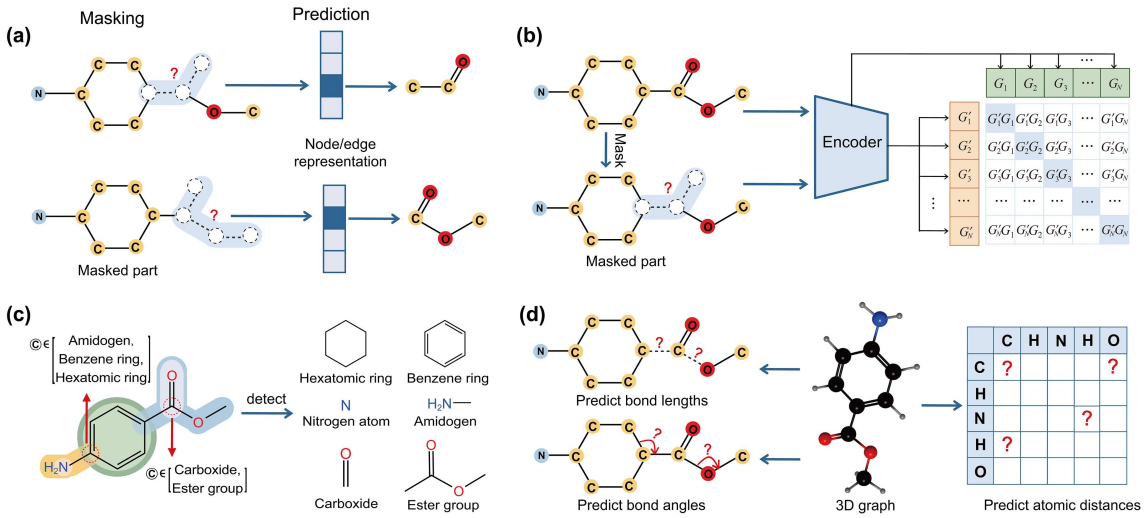

Figure 6 (Color online) Four pretraining strategies. (a) Mask-based pretraining. Predict information about the atom being masked. (b) Contrastive learning-based pretraining. Here, G represents the original graph, while G' refers to the enhanced graph. (c) Functional group-based pretraining. Identify information about functional groups within molecules. (d) Spatial structure-based pretraining. Predict bond lengths, bond angles, interatomic distances, etc.

5 Pretraining strategies

This section summarizes the pretraining tasks implemented in existing studies, categorically discussing the specific approaches employed for molecular representation learning. Molecular pretraining models can broadly be classified into four categories based on their design and objectives (Figure 6): mask-based pretraining, contrastive learning-based pretraining, functional group-based pretraining, and spatial structure-based pretraining. These categories reflect the diverse methodologies used to extract meaningful molecular representations, leveraging unique aspects of molecular structures and properties to enhance model performance across various downstream tasks.

5.1 Mask-based pretraining

In mask-based pretraining, the input molecular undergoes partial masking and the model is tasked with reconstructing this masked information. In sequence-based tasks, natural language processing-inspired approaches, such as ChemBERTa [60], mask portions of SMILES strings. Similarly, Chemformer [61], SMILES-BERT [59], FG-BERT [64], and PremuNet [74] employ comparable strategies. BioT5 [63] and BioT5+ [65] incorporate SELFIES and IUPAC to interpret the natural language descriptions and chemical structures of molecules, leveraging masked modeling to predict missing structural components.

In graph-based tasks, masking involves techniques such as atom masking, bond removal, and subgraph removal. For instance, GROVER [72] introduces a contextual property prediction task by masking certain molecular structures (local subgraphs) and predicting the properties of nodes and edges within these regions. MGSSL [23] employs only property masks for atoms and chemical bonds. MolCLR [66] extends the masking strategy to three levels: atoms, bonds, and subgraphs. It generates augmented molecular graphs through atom masking, bond deletion, and subgraph removal, employing a contrastive learning strategy to extract meaningful molecular representations. Similarly, Uni-Mol [75] incorporates atom species masking into its pretraining framework, using the prediction of masked atomic species as a core pretraining task, thereby enhancing the model's ability to learn detailed molecular features. KPGT, proposed by Li et al. [77], employs the LiGhT to capture complex structural patterns in molecular graphs for knowledge-guided learning. A pretraining strategy based on masked graphs is designed to effectively capture the structural and semantic knowledge of molecules.

ImageMol [18] utilizes molecular images for molecular characterization and introduces an image-based masking pretraining task by randomly masking regions of molecular images. This approach offers a novel perspective for mask-based pretraining.

5.2 Comparative learning-based pretraining

Contrastive learning has emerged as a prominent pretraining task in molecular representation learning, emphasizing the model's ability to distinguish between different molecules or their structural features by comparing positive and negative samples. This methodology is especially effective in scenarios without labeled data, significantly enhancing the generalization of molecular representations while preserving chemical semantics. Since contrastive learning requires both positive and negative samples, it is often integrated with masking-based pretraining techniques.

For example, MolCLR [66] employs a contrastive learning framework where each molecule is represented in two views: the original molecular graph and an augmented version. The augmented graph from the same molecule serves as a positive sample pair, while augmented graphs from different molecules act as negative sample pairs. Similarly, KANO [70] introduces a novel approach by constructing a knowledge graph based on the chemical elements of a molecule. It integrates knowledge graph embeddings and functional group hints to enrich molecular graph representations, which are subsequently used for contrastive learning pretraining. These strategies exemplify the versatility and effectiveness of contrastive learning in capturing intricate molecular relationships and structures. MolGT [80] employs implicit 3D geometric contrastive learning to align 2D and 3D molecular representations in feature space by maximizing mutual information, ensuring that the features of the same molecule remain consistent across both modalities.

5.3 Functional group-based pretraining

Functional group (FG)-based pretraining leverages molecular chemistry's a priori knowledge to improve a model's ability to comprehend molecular structures and properties by incorporating functional group and motif information. Functional groups are specific atom arrangements within a molecule that significantly influence its chemical behavior. Identifying these groups provides insights into the molecule's critical substructures and properties. Similar to functional groups, motifs are recurring substructural patterns in molecules that exhibit structural repetitions or functional correlations, offering broader coverage than functional groups. Both provide prior knowledge that enhances model performance and interpretability in molecular property prediction tasks.

GROVER [72] introduced a molecular-level motif identification task, focusing on detecting motifs such as functional groups and ring structures embedded in molecules. These motifs encode substantial domain knowledge essential for understanding molecular properties.

Zhang et al. [23] introduced MGSSL, the first method to employ a true functional group learning strategy rather than using motifs in a broader sense. They developed a molecular partitioning approach that leverages the inverse synthesis-based BRICS algorithm along with additional rules to regulate the size of the motif vocabulary. The molecular graph is subsequently transformed into a motif graph based on functional group division, thereby accounting for the positional relationships between functional groups during recognition.

Zang et al. [69] designed HMGNN and further proposed HiMol, a hierarchical molecular graph self-supervised learning framework. HiMol extracts graph representations of motif hierarchies from molecular graphs and achieves node-motif-graph hierarchical information representation by adding edges and constructing node-level, motif-level, and graph-level associations. This framework enables multi-layer self-supervised pre-training based on functional groups.

Li et al. [64] proposed FG-BERT, drawing inspiration from the BERT model [88]. FG-BERT implements masked modeling for functional groups within molecules to predict the masked molecular fragments, enhancing the model's ability to understand molecular structures.

Similarly, KANO [70] adopted a comparable approach, aiming to identify functional groups within a molecule. However, functional group classification in many methods often relies on dictionary-based lookups. To address this, Xie et al. [73] proposed a click-chemistry perceptual molecular partitioning approach and constructed a fragment-based molecular graph. Initially, molecular representations were derived through contrastive learning between SMILES and molecular graphs. These representations were then integrated with fragment-based molecular graph features, creating a more comprehensive molecular characterization.

However, their methodology limits functional group identification to the molecular level and does not refine the task to the atomic level, which could involve pinpointing the specific functional group associated

with individual atoms. This refinement could potentially unlock a more detailed understanding of molecular characteristics and enhance model precision. MolGT [80] captures key molecular motifs through contrastive learning, ensuring that atoms within the same motif share similar feature representations, while those in different motifs remain distinct. This enables atom-level motif feature extraction.

5.4 Molecular spatial structure-based pretraining

Spatial structure-based pretraining methods focus on utilizing the 3D spatial structure of molecule (such as bond lengths, bond angles, and atomic distances) to improve molecular representation learning. These methods are particularly relevant in applications like molecular property prediction and drug design, where spatial configurations play a critical role.

The most fundamental approach to incorporating a molecule's three-dimensional structure is by introducing its 3D coordinates and angles. GEM [34] proposed a novel approach by reconstructing the molecular graph into a bond angle graph, where bond angles are represented as edges and chemical bonds as nodes. Their work introduced three pretraining tasks: predicting bond lengths, predicting bond angles, and predicting inter-atomic distances. Notably, instead of directly predicting continuous atomic distances, GEM discretized the distances into 30 equal intervals, framing it as a multi-class classification task, which adds robustness to distance prediction.

GNS-TAT [67] introduces a 3D coordinate recovery task based on the 3D coordinates of molecules. The key idea is to add noise to atomic coordinates and predict this noise during pretraining, indirectly estimating the true atomic coordinates. This approach is both simple and innovative, serving as a pioneering effort in such pretraining tasks. Building on this, Liu et al. [68] proposed GeoSSL-DDM, refining the 3D coordinate denoising task through a fractional matching method leveraging SE(3)-equivariance. SE(3)-equivariance ensures that a neural network's representation remains unchanged under molecular rotations and translations, making it essential for 3D molecular learning tasks.

Subsequent approaches incorporated quantum chemical information to enhance molecular property prediction. Luo et al. [76] introduced Transformer-M, with its first pretraining task focused on a supervised learning objective to predict the energy gap of each molecule's HOMO-LUMO orbitals. Additionally, they employed a self-supervised learning objective called 3D positional noise reduction, which proved particularly effective. A 3D distance encoding matrix was used to aid training; however, the matrix is static and not updated during training, leaving room for improvement.

Building on the idea of incorporating 3D molecular information, Uni-Mol [75] introduced three innovative pretraining tasks: recovery of atom 3D coordinates, atom-atom pair distance prediction, and a masked pretraining task for atom classes. This method recovers 3D coordinates by denoising atomic positions. In Uni-Mol, the backbone architecture is flexible and can be replaced with any SE(3)-equivariant model capable of processing 3D positions as inputs and outputs. Additionally, the interatomic distance representation is dynamically updated during training, enhancing its adaptability and effectiveness. Expanding on Uni-Mol, Uni-Mol+ [79] introduced QC property prediction tasks, which include predicting energies, polarizabilities, and molecular orbitals. These properties are directly tied to the chemical structure and reveal crucial information about a molecule's electronic structure, reactivity, and stability. By integrating QC properties into pretraining, Uni-Mol+ enables deeper and more precise learning of molecular characteristics, enhancing its applicability to complex chemical analyses.

Mol-AE, proposed by Yang et al. [78], builds upon the concept of Uni-Mol by introducing a novel training objective called the 3D Cloze Test. This approach enables the model to better capture the spatial relationships among atoms in real molecular structures.

Ni et al. [71] further enhanced the denoising pretraining task with the Frad framework. They introduced a mixture of chemical awareness noise (CAN) and coordinate Gaussian noise (CGN) to generate noisy molecular conformations. During pretraining, the model predicts the CGN noise, effectively decoupling noise design from the constraints of force-learning equivalence. This customizable noise design allows the incorporation of chemical priors, significantly improving the performance of molecular distribution models.

PremuNet [74] designs three pretraining tasks to facilitate the interaction between 2D and 3D molecular information: (1) masking all atoms and reconstructing them using 3D coordinates; (2) masking all 3D coordinates and reconstructing them using atomic information; and (3) independently masking atoms and their coordinates, then reconstructing them using the features of unmasked atoms and 3D coordinates. These tasks complement each other-the first enables the model to extract 3D information from 2D data,

the second allows it to extract 2D information from 3D data, and the third integrates both modalities, enhancing the model's multimodal learning capability.

5.5 Other pretraining strategies

ImageMol [18] introduces an innovative approach by transforming molecules into molecular images and leveraging computer vision for molecular representation learning. Its pretraining strategy includes tasks such as masked comparison learning on images, molecular image reconstruction, and the multi-granularity chemical clusters classification (MG3C) task, which ensures pretraining consistency. Additionally, ImageMol incorporates rationality-focused tasks, including molecular rationality discrimination and jigsaw puzzle prediction, to align the structural information of molecular images with established chemical principles. While the method demonstrates promising performance, the use of images for molecular representation learning requires further exploration.

VideoMol [20] extends molecular representation learning to a dynamic context, incorporating three pretraining tasks: video-aware, orientation-aware, and chemistry-aware pretraining. Video-aware pretraining trains the model to distinguish between different molecular videos, such as identifying whether two frames belong to the same video. Orientation-aware pretraining focuses on recognizing spatial relationships, such as angular differences between frames. Chemistry-aware pretraining extracts physicochemical information from molecular videos. Together, these tasks enable dynamic and physicochemical perception of molecules, offering a novel perspective for molecular pretraining models.

PremuNet [74] employs an AutoEncoder framework for processing SMILES representations. It first tags the original SMILES string, and then encodes it using a multilayer encoder to generate a feature matrix. This matrix is subsequently decoded by a multilayer decoder, minimizing the difference between the input and output. This pretraining approach enables the transformer encoder to efficiently extract meaningful features from SMILES strings.

MolGT [80] introduces knowledge-guided prototypical clustering (KGPC) for layer-level pretraining from both 2D and 3D perspectives, leveraging MACCS and USRCAT molecular fingerprints as prior knowledge.

6 Performance comparison

We identified the best-performing components of the models discussed in this paper and evaluated their performance on a molecular property prediction dataset. For the models included in the comparison, we referenced data from their original publications and assessed their performance using both scaffold split and random scaffold split divisions (Tables 4–7). Some method performance is missing and therefore not counted in the table.

Synthesizing the various approaches reveals several key observations. (1) Most methods adopt scaffold split as the primary evaluation criterion, with only a few considering both scaffold split and random scaffold split scenarios. (2) Performance in the random scaffold split scenario consistently exceeds that in the scaffold split scenario, highlighting the increased challenge posed by molecular property prediction under scaffold split conditions. (3) No single method achieves optimal performance across all tasks, underscoring the need for task-specific optimization. (4) Methods incorporating additional information, such as 3D spatial structure or functional groups, generally outperform those relying solely on 2D molecular graphs, demonstrating the value of enriched molecular representations.

7 Applications

This section explores various applications of predictive modeling based on molecular properties.

7.1 Target-based drug discovery

By leveraging structural information of specific biological targets, high-affinity small-molecule inhibitors can be identified through a combination of computational simulations and experimental screening. For instance, hematopoietic progenitor kinase 1 (HPK1) and fibroblast growth factor receptor 1 (FGFR1) are implicated in various cancer types and have been extensively studied for antitumor therapy [89–92].

Table 4 ROC-AUC (%) performance of scaffold split scenes for some excellent methods on classification tasks. All results are reported as mean. Higher is better.

	BACE	BBBP	ClinTox	ToxCast	Tox21	SIDER	HIV	MUV
MolCLR	81.9	71.6	91.9	–	75.0	59.9	78.3	79.7
GROVER	82.2	71.8	84.3	63.5	76.5	63.7	78.6	76.9
MGSSL	79.1	69.7	80.7	64.1	76.5	61.8	78.8	78.7
ImageMol	83.9	73.9	85.1	65.9	77.3	67.7	79.7	82.5
GEM	85.6	72.4	90.1	69.2	78.1	67.2	80.6	81.7
Uni-Mol	85.7	72.9	91.9	69.6	79.6	65.9	80.8	82.1
MOL-AE	84.1	72.0	87.8	69.0	80.0	67.0	80.6	81.6
KPGT	85.5	–	94.6	74.6	84.8	74.6	–	–
HiMol	84.6	73.2	80.8	66.3	76.2	62.5	–	–
BioT5	89.4	77.7	95.4	–	77.9	73.2	81.0	–
MolGT	84.5	73.7	88.9	66.4	75.8	65.4	79.3	78.9
PremuNet	84.3	73.3	99.2	–	74.0	62.6	–	–

Table 5 Performance of the regression task in the scaffold split scenario. All results are reported as mean. Higher is better.

	RMSE			MAE		
	FreeSolv	ESOL	Lipophilicity	QM7	QM8	QM9
MolCLR	2.47	1.21	0.69	144.4	0.0359	0.01488
GROVER	2.48	0.99	0.66	92.0	0.0224	0.00986
ImageMol	2.02	0.97	0.72	116.4	0.0241	0.02061
GEM	1.877	0.798	0.660	58.9	0.0171	0.00746
Uni-Mol	1.480	0.788	0.603	41.8	0.0156	0.00467
MOL-AE	1.448	0.830	0.607	53.8	0.0161	0.00530
KPGT	2.121	0.803	0.600	–	–	–
HiMol	2.283	0.833	0.708	91.5	0.0199	–
MolGT	–	0.839	0.788	–	–	–
PremuNet	1.858	0.730	–	–	–	–

Table 6 ROC-AUC (%) performance of random scaffold split scenes for some excellent methods on classification tasks. All results are reported as mean. Higher is better.

	BACE	BBBP	ClinTox	ToxCast	Tox21	SIDER	HIV	MUV
MolCLR	89.0	73.6	93.2	–	79.8	68.0	80.6	88.6
GROVER	92.3	94.0	95.6	74.1	84.0	69.1	–	–
KANO	93.1	96.0	94.4	73.2	83.7	65.2	85.1	83.7
ImageMol	93.9	95.2	97.5	75.2	84.7	70.8	–	–

Table 7 Performance of the regression task in the random scaffold split scenario. All results are reported as mean. Higher is better.

	RMSE			MAE		
	FreeSolv	ESOL	Lipophilicity	QM7	QM8	QM9
MolCLR	2.20	1.11	0.65	87.2	0.0174	–
GROVER	1.366	0.730	0.556	72.1	–	–
KANO	1.142	0.670	0.566	56.4	0.0123	0.00320
ImageMol	1.149	0.690	0.625	65.9	–	–

KPGT enables efficient screening of HPK1 and FGFR1 inhibitors, demonstrating strong performance in both structural and temporal partitioning test sets.

7.2 Activity cliff analysis

Activity cliffs refer to structurally similar molecules that exhibit significant differences in biological activity. Conventional molecular property prediction models typically rely on structural similarity principles; however, activity cliff molecules can differ in activity by severalfold or even hundreds of times despite high structural similarity. This can lead to incorrect learning of non-deterministic substructures. Designing molecular pretraining models to capture complex structural features can help identify key determinants

that differentiate structurally similar molecules with varying activities, thereby improving recognition accuracy.

For instance, van Tilborg et al. [93] constructed the MoleculeACE dataset to evaluate the performance of various deep learning methods on activity cliff analysis. Additionally, KPGT [77] further explored activity cliff recognition using the CYP3A4 dataset. Results demonstrated that KPGT could accurately predict biological activity differences between activity cliff molecule pairs with high fingerprint similarity, highlighting its ability to sensitively capture subtle variations in molecular activity.

8 Future outlooks

Although pre-trained models are now widely used in the field of molecular characterization, there are still a number of issues that must be addressed. We summarize some of the current issues and discuss possible directions for improvement.

8.1 Improved interactions between different molecular modalities

Existing molecular descriptors, such as SMILES, molecular fingerprints, 2D molecular graphs, 3D molecular graphs, and molecular images, encompass diverse modalities, including sequences, planar graphs, and 3D spatial conformations. Current methods have explored multimodal knowledge integration through various approaches, such as using multimodal information as model input or incorporating multimodal pretraining tasks. However, more efficient strategies are required to optimize the learning of multimodal information. Multimodal pretraining frameworks like CLIP [94] and ALBEF [95] represent promising directions. Furthermore, adopting knowledge distillation architectures-where complex models (e.g., molecular image-based or video-based models) serve as teacher models to guide the training of simpler models (e.g., molecular graph-based models)-could significantly enhance the performance of molecular pretraining models.

8.2 Optimizing the impact of pretraining tasks

Pretraining tasks for molecular models can encompass various aspects, including molecular properties, functional groups, and 3D conformations. The importance of these tasks varies in influencing model learning and should be carefully considered during training. A potential solution is to assign different weights to pretraining tasks, enabling the development of an optimized training strategy for molecular representation models.

8.3 Explainability of pretraining tasks

Deep learning models are often perceived as "black boxes" due to their limited interpretability, which hinders their broader application in real-world scenarios across various domains. This challenge extends to pretraining tasks, where methods predict features such as chemical bond lengths and angles without clarifying how these predictions enable the model to acquire new knowledge. For example, 3D bond angle prediction allows models to reason the 3D structure given 2D graphs, thus enabling learning information not presented in the original input graph. Future research should prioritize interpretability, focusing on approaches that provide credible insights to support applications like drug development.

9 Conclusion

This review provides an overview of sequence-based, 2D-based, and 3D-based molecular representations, with a particular emphasis on molecular pretraining models. We discuss various model architectures and pretraining strategies, using the molecular property prediction task as a case study to introduce mature datasets and evaluation criteria while comparing the performance of state-of-the-art methods. Additionally, we examine the limitations of existing molecular pretraining models and propose potential directions for improvement. Looking ahead, we aim to encourage the development and application of high-performing models in molecular characterization to enhance molecular property prediction and facilitate drug discovery.

Acknowledgements This work was supported by National Science and Technology Major Project of China (Grant No. 2023ZD01-20903), National Natural Science Foundation of China (Grant No. 62322112), Science and Technology Development Fund of Macao (Grant No. 0133/2024/RIB2), and National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (Grant No. RS-2024-00344752).

References

1 Battaglia P W, Hamrick J B, Bapst V, et al. Relational inductive biases, deep learning, and graph networks. 2018. ArXiv:1806.01261
2 Chen L, Yu L, Gao L, et al. Potent antibiotic design via guided search from antibacterial activity evaluations. Bioinformatics, 2023, 39: btad059
3 Wang Y Z, Zhai Y X, Ding Y J, et al. SBSM-Pro: support bio-sequence machine for proteins. Sci China Inf Sci, 2024, 67: 212106
4 Zou Q, Xing P, Wei L, et al. Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. RNA, 2019, 25: 205–218
5 Zulfiqar H, Guo Z, Ahmad R M, et al. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. Front Med, 2024, 10: 1291352
6 Ai C, Yang H, Liu X, et al. MTMol-GPT: de novo multi-target molecular generation with transformer-based generative adversarial imitation learning. PLoS Comput Biol, 2024, 20: e1012229
7 Ghasemi F, Mehridehnavi A, Pérez-Garrido A, et al. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov Today, 2018, 23: 1784–1790
8 Huang Z, Chen S, Yu L. Predicting new drug indications based on double variational autoencoders. Comput Biol Med, 2023, 164: 107261
9 Zhu H, Hao H, Yu L. Identification of microbe-disease signed associations via multi-scale variational graph autoencoder based on signed message propagation. BMC Biol, 2024, 22: 172
10 Joshi M, Singh B K. Deep learning techniques for brain lesion classification using various MRI (from 2010 to 2022): review and challenges. Medinformatics, 2024. https://ojs.bonviewpress.com/index.php/MEDIN/article/view/1686
11 Ai C, Yang H, Ding Y, et al. Low rank matrix factorization algorithm based on multi-graph regularization for detecting drug-disease association. IEEE ACM Trans Comput Biol Bioinf, 2023, 20: 3033–3043
12 Li T, Ren X, Luo X, et al. A foundation model identifies broad-spectrum antimicrobial peptides against drug-resistant bacterial infection. Nat Commun, 2024, 15: 7538
13 Tao W, Liu Y, Lin X, et al. Prediction of multi-relational drug-gene interaction via dynamic hypergraph contrastive learning. Brief BioInf, 2023, 24: bbad371
14 Liu M, Li C, Chen R, et al. Geometric deep learning for drug discovery. Expert Syst Appl, 2024, 240: 122498
15 Wang Z, Chen Y, Shang Y, et al. MultiCycPermea: accurate and interpretable prediction of cyclic peptide permeability using a multimodal image-sequence model. BMC Biol, 2025, 23: 63
16 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci, 1988, 28: 31–36
17 Yang Y, Gao D, Xie X, et al. DeepIDC: a prediction framework of injectable drug combination based on heterogeneous information and deep learning. Clin Pharmacokinet, 2022, 61: 1749–1759
18 Zeng X, Xiang H, Yu L, et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. Nat Mach Intell, 2022, 4: 1004–1016
19 Wang Z, Chen Y, Ma P, et al. Image-based generation for molecule design with SketchMol. Nat Mach Intell, 2025, 7: 244–255
20 Xiang H, Zeng L, Hou L, et al. A molecular video-derived foundation model for scientific drug discovery. Nat Commun, 2024, 15: 9696
21 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2020, 32: 4–24
22 Yang X, Zhao X, Shen Z. A generalizable anomaly detection method in dynamic graphs. 2024. ArXiv:2412.16447
23 Zhang Z, Liu Q, Wang H, et al. Motif-based graph self-supervised learning for molecular property prediction. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 15870–15882
24 Li P, Wang J, Qiao Y, et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. Brief BioInf, 2021, 22: bbab109
25 Krenn M, Häse F, Nigam A K, et al. Self-referencing embedded strings (selfies): a 100% robust molecular string representation. Mach Learn Sci Technol, 2020, 1: 045024
26 Klinger R, Kolářik C, Fluck J, et al. Detection of IUPAC and IUPAC-like chemical names. Bioinformatics, 2008, 24: 268–276
27 Duvenaud D K, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints. In: Proceedings of Advances in Neural Information Processing Systems, 2015
28 Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res, 2019, 47: e127
29 Li H L, Pang Y H, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. Nucleic Acids Res, 2021, 49: e129
30 Liu X W, Shi T Y, Gao D, et al. iPADD: a computational tool for predicting potential antidiabetic drugs using machine learning algorithms. J Chem Inf Model, 2023, 63: 4960–4969
31 Durant J L, Leland B A, Henry D R, et al. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci,

2002, 42: 1273–1280
32 Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model, 2010, 50: 742–754
33 Xia J, Zhu Y, Du Y, et al. A systematic survey of molecular pre-trained models. 2022. ArXiv:2210.16484
34 Fang X, Liu L, Lei J, et al. Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell, 2022, 4: 127–134
35 Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res, 2019, 47: D1102–D1109
36 Sterling T, Irwin J J. ZINC 15—ligand discovery for everyone. J Chem Inf Model, 2015, 55: 2324–2337
37 Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. Nucleic Acids Res, 2017, 45: D945–D954
38 Hu W, Fey M, Ren H, et al. OGB-LSC: a large-scale challenge for machine learning on graphs. 2021. ArXiv:2103.09430
39 Axelrod S, Gómez-Bombarelli R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. Sci Data, 2022, 9: 185
40 Isert C, Atz K, Jiménez-Luna J, et al. QMugs, quantum mechanical properties of drug-like molecules. Sci Data, 2022, 9: 273
41 Xu Z, Luo Y, Zhang X, et al. Molecule3D: a benchmark for predicting 3D geometries from molecular graphs. 2021. ArXiv:2110.01717
42 Wu Z, Ramsundar B, Feinberg E N, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci, 2018, 9: 513–530
43 Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn, 1997, 30: 1145–1159
44 Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)? — Arguments against avoiding RMSE in the literature. Geosci Model Dev, 2014, 7: 1247–1250
45 Subramanian G, Ramsundar B, Pande V, et al. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. J Chem Inf Model, 2016, 56: 1936–1949
46 Martins I F, Teixeira A L, Pinheiro L, et al. A Bayesian approach to in silico blood-brain barrier penetration modeling. J Chem Inf Model, 2012, 52: 1686–1697
47 Gayvert K M, Madhukar N S, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. Cell Chem Biol, 2016, 23: 1294–1301
48 Huang R, Xia M, Nguyen D T, et al. Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci, 2016, 3: 85
49 Richard A M, Judson R S, Houck K A, et al. ToxCast chemical landscape: paving the road to 21st century toxicology. Chem Res Toxicol, 2016, 29: 1225–1251
50 Kuhn M, Letunic I, Jensen L J, et al. The SIDER database of drugs and side effects. Nucleic Acids Res, 2016, 44: 1075–1079
51 Rohrer S G, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. J Chem Inf Model, 2009, 49: 169–184
52 Riesen K, Bunke H. IAM graph database repository for graph based pattern recognition and machine learning. In: Proceedings of Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2008. 287–297
53 Mobley D L, Guthrie J P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des, 2014, 28: 711–720
54 Delaney J S. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci, 2004, 44: 1000–1005
55 Gaulton A, Bellis L J, Bento A P, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res, 2012, 40: D1100–D1107
56 Blum L C, Reymond J L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc, 2009, 131: 8732–8733
57 Ramakrishnan R, Hartmann M, Tapavicza E, et al. Electronic spectra from TDDFT and machine learning in chemical space. J Chem Phys, 2015, 143: 084111
58 Ruddigkeit L, van Deursen R, Blum L C, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model, 2012, 52: 2864–2875
59 Wang S, Guo Y, Wang Y, et al. 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, 2019. 429–436
60 Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. 2020. ArXiv:2010.09885
61 Irwin R, Dimitriadis S, He J, et al. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn-Sci Technol, 2022, 3: 015022
62 Ahmad W, Simon E, Chithrananda S, et al. Chemberta-2: towards chemical foundation models. 2022. ArXiv:2209.01712
63 Pei Q, Zhang W, Zhu J, et al. Biot5: enriching cross-modal integration in biology with chemical knowledge and natural language associations. 2023. ArXiv:2310.07276
64 Li B, Lin M, Chen T, et al. FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. Brief BioInf, 2023, 24: bbad398
65 Pei Q, Wu L, Gao K, et al. Biot5+: towards generalized biological understanding with IUPAC integration and multi-task

tuning. 2024. ArXiv:2402.17810
66 Wang Y, Wang J, Cao Z, et al. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell, 2022, 4: 279–287
67 Zaidi S, Schaarschmidt M, Martens J, et al. Pre-training via denoising for molecular property prediction. 2022. ArXiv:2206.00133
68 Liu S, Guo H, Tang J. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. 2022. ArXiv:2206.13602
69 Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction. Commun Chem, 2023, 6: 34
70 Fang Y, Zhang Q, Zhang N, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell, 2023, 5: 542–553
71 Ni Y, Feng S, Hong X, et al. Pre-training with fractional denoising to enhance molecular property prediction. Nat Mach Intell, 2024, 6: 1169–1178
72 Rong Y, Bian Y, Xu T, et al. Self-supervised graph transformer on large-scale molecular data. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 12559–12571
73 Xie A, Zhang Z, Guan J, et al. Self-supervised learning with chemistry-aware fragmentation for effective molecular property prediction. Brief BioInf, 2023, 24: bbad296
74 Zhang H, Wu J, Liu S, et al. A pre-trained multi-representation fusion network for molecular property prediction. Inf Fusion, 2024, 103: 102092
75 Zhou G, Gao Z, Ding Q, et al. Uni-Mol: a universal 3D molecular representation learning framework. In: Proceedings of the 11th International Conference on Learning Representations, 2023
76 Luo S, Chen T, Xu Y, et al. One transformer can understand both 2D & 3D molecular data. In: Proceedings of the 11th International Conference on Learning Representations, 2023
77 Li H, Zhang R, Min Y, et al. A knowledge-guided pre-training framework for improving molecular representation learning. Nat Commun, 2023, 14: 7568
78 Yang J, Zheng K, Long S, et al. Mol-AE: auto-encoder based molecular representation learning with 3D cloze test objective. In: Proceedings of the 41st International Conference on Machine Learning, 2024. 56793–56811
79 Lu S, Gao Z, He D, et al. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. Nat Commun, 2024, 15: 7104
80 Chen R, Li C, Wang L, et al. Pretraining graph transformer for molecular representation with fusion of multimodal information. Inf Fusion, 2025, 115: 102784
81 Lewis M. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2019. ArXiv:1910.13461
82 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2016. ArXiv:1609.02907
83 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2017. ArXiv:1710.10903
84 Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks? 2018. ArXiv:1810.00826
85 Gilmer J, Schoenholz S S, Riley P F, et al. Neural message passing for quantum chemistry. In: Proceedings of International Conference on Machine Learning, 2017. 1263–1272
86 Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of Advances in Neural Information Processing Systems, 2017
87 Song Y, Zheng S, Niu Z, et al. Communicative representation learning on attributed molecular graphs. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020. 2831–2838
88 Tran H V, Nguyen Q H. iAnt: combination of convolutional neural network and random forest models using PSSM and BERT features to identify antioxidant proteins. Curr Bioinform, 2022, 17: 184–195
89 Shui J W, Boomer J S, Han J, et al. Hematopoietic progenitor kinase 1 negatively regulates T cell receptor signaling and T cell-mediated immune responses. Nat Immunol, 2007, 8: 84–91
90 Si J, Shi X, Sun S, et al. Hematopoietic progenitor kinase1 (HPK1) mediates T cell dysfunction and is a druggable target for T cell-based immunotherapies. Cancer Cell, 2020, 38: 551–566.e11
91 Acevedo V D, Gangula R D, Freeman K W, et al. Inducible FGFR-1 activation leads to irreversible prostate adenocarcinoma and an epithelial-to-mesenchymal transition. Cancer Cell, 2007, 12: 559–571
92 Nguyen P T, Tsunematsu T, Yanagisawa S, et al. The FGFR1 inhibitor PD173074 induces mesenchymal-epithelial transition through the transcription factor AP-1. Br J Cancer, 2013, 109: 2248–2258
93 van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model, 2022, 62: 5938–5951
94 Li Y, Liang F, Zhao L, et al. Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm. 2021. ArXiv:2110.05208
95 Li J, Selvaraju R, Gotmare A, et al. Align before fuse: vision and language representation learning with momentum distillation. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 9694–9705