Chinese Journal of Natural Medicines

Advances in small molecule representations and AI-driven drug research: bridging the gap between theory and application

Junxi Liu, Shan Chang, Qingtian Deng, Yulian Ding, Yi Pan

Citation: Junxi Liu, Shan Chang, Qingtian Deng, Yulian Ding, Yi Pan, Advances in small molecule representations and AI-driven drug research: bridging the gap between theory and application, Chinese Journal of Natural Medicines, 2025, 23(11), 1391-1408. doi: 10.1016/S1875-5364(25)60946-0.

View online: https://doi.org/10.1016/S1875-5364(25)60946-0

Chinese Journal of Natural Medicines 23 (2025) 1391-1408

Contents lists available at ScienceDirect

Chinese Journal of Natural Medicines

journal homepage: www.cjnmcpu.com/

Review

Advances in small molecule representations and AI-driven drug research: bridging the gap between theory and application

Junxi Liu[a,b], Shan Chang[c], Qingtian Deng[b], Yulian Ding[d,*], Yi Pan[b,e,*]

[a] *Shenzhen University of Advanced Technology, Southern University of Science and Technology, Shenzhen 518055, China*
[b] *Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen 518107, China*
[c] *Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou 213001, China*
[d] *Central for High Performance Computing, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[e] *Shenzhen Key Laboratory of Intelligent Bioinformatics, Shenzhen Institute of Advanced Technology, Shenzhen 518055, China*

ARTICLE INFO

Article history:
Received 1 February 2025
Revised 13 March 2025
Accepted 29 April 2025
Available online 20 November 2025

Keywords:
Small molecular representation
Drug-target interaction prediction
Drug-target affinity prediction
Drug property prediction
De novo drug generation
Traditional Chinese medicine

ABSTRACT

Artificial intelligence (AI) researchers and cheminformatics specialists strive to identify effective drug precursors while optimizing costs and accelerating development processes. Digital molecular representation plays a crucial role in achieving this objective by making molecules machine-readable, thereby enhancing the accuracy of molecular prediction tasks and facilitating evidence-based decision making. This study presents a comprehensive review of small molecular representations and AI-driven drug discovery downstream tasks utilizing these representations. The research methodology begins with the compilation of small molecule databases, followed by an analysis of fundamental molecular representations and the models that learn these representations from initial forms, capturing patterns and salient features across extensive chemical spaces. The study then examines various drug discovery downstream tasks, including drug-target interaction (DTI) prediction, drug-target affinity (DTA) prediction, drug property (DP) prediction, and drug generation, all based on learned representations. The analysis concludes by highlighting challenges and opportunities associated with machine learning (ML) methods for molecular representation and improving downstream task performance. Additionally, the representation of small molecules and AI-based downstream tasks demonstrates significant potential in identifying traditional Chinese medicine (TCM) medicinal substances and facilitating TCM target discovery.

1. Introduction

The traditional approach to drug discovery faces significant challenges, characterized by substantial resource requirements and low success rates [1]. The progression from initial concept to a market-approved drug requires extensive resources and commitment. According to the Tufts Center for the Study of Drug Development, developing a new molecular entity requires approximately 2.6 billion US dollars [2]. This substantial cost encompasses multiple components, including high-throughput screening of compounds, comprehensive preclinical and clinical trials, and advanced research infrastructure requirements. The development and approval process typically spans 10–15 years [3]. The success rate in conventional drug development remains notably low. An analysis of over 10 000 drug development projects revealed that merely 12% of drugs entering Phase I clinical trials achieve regulatory approval [4]. A previous study indicated that among candidate drugs, less than 0.1% advanced from preclinical testing to human trials, with only 20% of these candidates successfully entering the market [5]. This limited success rate stems from the vast and discrete nature of chemical molecular space. The potential structures of drug-like compounds range between 10^{23} and 10^{60}, with approximately 10^{8} having therapeutic relevance. Traditional methodologies such as high-throughput screening demonstrate limited efficiency. However, big data and advanced computing capabilities offer new opportunities to overcome these limitations.

The accumulation of drug-related clinical data has led to the establishment of comprehensive databases, including DrugBank 6.0 [6] and PubChem [7]. DrugBank 6.0 functions as a comprehensive repository for drug-related information, drug targets, and pharmaceutical data. PubChem represents a widely-used chemical information dataset containing chemical structures, identifiers, and chemical and physical properties. As cheminformatics knowledge expands and artificial intelligence (AI) technologies advance, AI-driven drug research has emerged as a novel paradigm. Various machine learning (ML) techniques demonstrate effectiveness in this domain, with their computational capabilities efficiently managing complex tasks. Basic ML methods, such as random forest (RF), effectively perform classification tasks using human-designed features [8,9]. Convolutional neural networks (CNNs) extract features from matrices, while recurrent neural networks (RNNs), particularly long short-term memory networks (LSTMs), excel in processing sequential information [10,11]. Contemporary large language models (LLMs), including GPT and Bert, demon-

* Corresponding author.
E-mail addresses: yl.ding2@siat.ac.cn (Y. Ding); yi.pan@siat.ac.cn (Y. Pan)

https://doi.org/10.1016/S1875-5364(25)60946-0

strate versatility across tasks and provide novel approaches for various stages of drug development [12, 13]. AI-driven drug discovery presents opportunities to transform the pharmaceutical industry by optimizing development processes, reducing costs, and enhancing the identification of safe and effective drugs [14, 15].

While AI technology significantly enhances drug development, the effectiveness of AI-driven drug research depends substantially on molecular digital description and representation learning from this description. Digital molecular representation facilitates exploration of chemical space and reveals molecular patterns and relationships. Converting molecular structures into machine-readable formats enables AI analysis and prediction of molecular behavior with enhanced precision and efficiency. Computer-based molecular representation differs from human perception, and molecular description has evolved across disciplines. In 1919, researchers initially utilized "common names" for molecular identification, such as the International Union of Pure and Applied Chemistry (IUPAC), which standardized chemical nomenclature [16]. This nomenclature requires extensive knowledge of chemical moieties, naming conventions, and syntax. Subsequently, alternative molecular representation formats emerged, better suited for computational analysis, such as the simplified molecular input line entry system (SMILES). Currently, numerous molecular digital formats exist, with the open-source Open Babel program supporting 146 distinct basic formats. These digital formats are generally categorized into string-based and graph-based representations. Various ML methods utilize these digital formats to extract molecular feature vectors, generating representations applicable to different downstream tasks.

The SMILES representation stands as the most widely used string-based representation, encoding molecules as sequences of letters and symbols. It denotes atoms by their atomic symbols, bonds by distinct symbols, encloses branches in parentheses, and utilizes atom-number combinations for closed rings [17]. While the SMILES format offers both human and machine readability, multiple synonymous representations can exist for the same molecule. Subsequently, extensions and alternatives to SMILES were developed, including SMILES arbitrary target specification (SMARTS) and SELFIES. Molecular fingerprints represent another string-based approach, converting molecules into numerical or vector representations through dictionary or path-based methods. Although string-based representations facilitate machine readability, they typically lack specific spatial atomic information beyond connectivity. Graph-based representations emerged to address this limitation, depicting molecules as molecular graphs where edges and nodes correspond to bonds and atoms, respectively, with some variants incorporating 3D information and 2D molecular imagery [18]. Furthermore, molecules can be represented within virtual 3D grids. Precise and meaningful molecular representation proves essential for enabling AI algorithms to perform various drug discovery-related prediction tasks [19]. The comprehensiveness of molecular representation directly influences the completeness of extractable feature information. An optimal molecular representation should preserve task-relevant features to enhance subsequent network training performance.

After successfully representing small molecules in a form suitable for computer operations, researchers can undertake a series of drug development tasks with the assistance of AI. The AlphaFold 3 proposed by Google DeepMind in 2024 significantly accelerates AI-driven drug research by providing highly accurate predictions of protein-ligand interactions and molecular structures [20]. Its advanced AI capabilities allow for rapid identification of potential drug candidates and optimization of lead compounds, reducing both time and costs associated with traditional experimental methods. Although it demonstrates excellent performance in predicting biological molecular structures, it faces limitations in capturing the dynamic conception of protein-ligand binding and cannot predict binding strength and kinetic simulation parameters. Comparative analyses of AlphaFold3-predicted GPCR structures and experimentally determined counterparts demonstrate that, while AlphaFold3 achieves superior accuracy over AlphaFold2 in modeling the overall GPCR backbone conformation, notable deviations persist in predicted ligand-binding configurations, particularly for ions, peptides, and protein ligands [21]. Consequently, AI-based drug-related downstream tasks remain essential. For instance, *de novo* small molecule generation (MG) involves creating entirely new molecules from scratch using AI algorithms. These algorithms can learn from existing molecular structures and generate new compounds with desired properties. For example, deep generative models can learn data distributions and generate new samples, potentially leading to the discovery of new drug molecules [22]. Following the identification of novel drug candidates, artificial intelligence can be applied to predict their potential interactions with specific biological targets, a process referred to as drug–target interaction (DTI) prediction. DTI prediction is integral to both drug repositioning and *de novo* virtual screening. By systematically mining large-scale DTI datasets, AI models can identify potential drug–target associations, thereby reducing the need for labor-intensive and costly experimental validation [23]. Beyond identifying interactions, AI can also quantify their strength through drug–target affinity (DTA) prediction, which is critical for assessing binding efficiency and, consequently, therapeutic efficacy. These predictions draw on diverse data sources, including chemical structures and biological profiles [24]. For example, graph neural networks have been applied to capture complex relational patterns between drugs and targets, enabling more accurate affinity estimation [25]. AI further facilitates the prediction of a wide spectrum of drug properties (DP), including pharmacokinetics—encompassing absorption, distribution, metabolism, excretion, and toxicity—and pharmacodynamics, which characterize the biochemical and physiological effects of drugs on the body [26]. Early-stage DP prediction supports the evaluation of drug safety and efficacy, offering substantial savings in both time and resources during development [27, 28]. The integration of Traditional Chinese Medicine (TCM) with AI has emerged as a prominent research frontier [29, 30]. Similar downstream tasks in AI-based TCM research include the identification of active compounds, elucidation of their molecular targets, and prediction of their pharmacological properties. The effectiveness of these applications—and of AI-driven drug discovery more broadly—ultimately depends on the quality and representational power of the underlying drug descriptors.

This study examines critical aspects of small-molecule representation in the context of AI-driven drug research. Section 2 introduces widely used datasets pertinent to small molecules. Section 3 describes foundational digital molecular representations—such as SMILES notation, molecular fingerprints, molecular graphs, 2D images, 3D grids, and hybrid formats—and further discusses learned representations derived from these foundational forms. Section 4 reviews the application of these representations to key downstream tasks, including DTA prediction, DTI prediction, DP prediction, and de novo small MG. Finally, Section 5 synthesizes the main findings and outlines current challenges and future opportunities for advancing molecular representation techniques and enhancing the performance of diverse AI-driven downstream tasks.

A schematic overview of the study's structure and key components is presented in Fig. 1.

2. Datasets

AI, particularly deep neural networks, has been widely implemented across various stages of drug development. In AI-as-

sisted drug design, high-quality small-molecule data serves as a fundamental component. A significant volume of such data is essential for effectively training deep neural networks, as data quality directly influences AI algorithm performance. Researchers have compiled numerous publicly accessible small molecule datasets specifically designed for relevant training purposes. This section presents and analyzes several commonly utilized small-molecule datasets in related tasks.

2.1. DGIdb 5.0

The DGIdb 5.0 is a publicly accessible dataset that emphasizes drug-gene interactions. Clinicians and researchers frequently utilize it for drug development studies. The dataset encompasses genes, gene products, and their corresponding drug-gene interaction records. It offers drug data capable of interacting with genes or gene products of interest. However, the database exhibits limited coverage of non-coding RNA or epigenetic targets. Furthermore, diverse annotation criteria and evidence grading across sources may lead to conflicting or redundant information [31].

2.2. DrugBank 6.0

The DrugBank 6.0 is a publicly accessible dataset that functions as a comprehensive resource for drug-related information, drug targets, and associated pharmaceutical data. It finds extensive application in biomedical research and clinical settings. The dataset contains detailed information on drugs and drug targets, encompassing chemistry, pharmacology, pharmacokinetics, drug interactions, and drug-food interactions. Nevertheless, it shows limitations in niche drug classes and emerging therapies such as gene-based treatments. The clinical data lacks patient-level trial details, and commercial access fees may limit academic usage, while real-time updates remain unavailable [6].

2.3. IUPHAR/BPS in 2024

The IUPHAR/BPS in 2024 excels in expert-curated, authoritative coverage of over 3 000 protein targets and 12 000 ligands, including drugs, peptides, and antibodies. Its strengths lie in structured pharmacological data integration (e.g., binding kinetics, signaling pathways), regular updates (four releases in 2024),

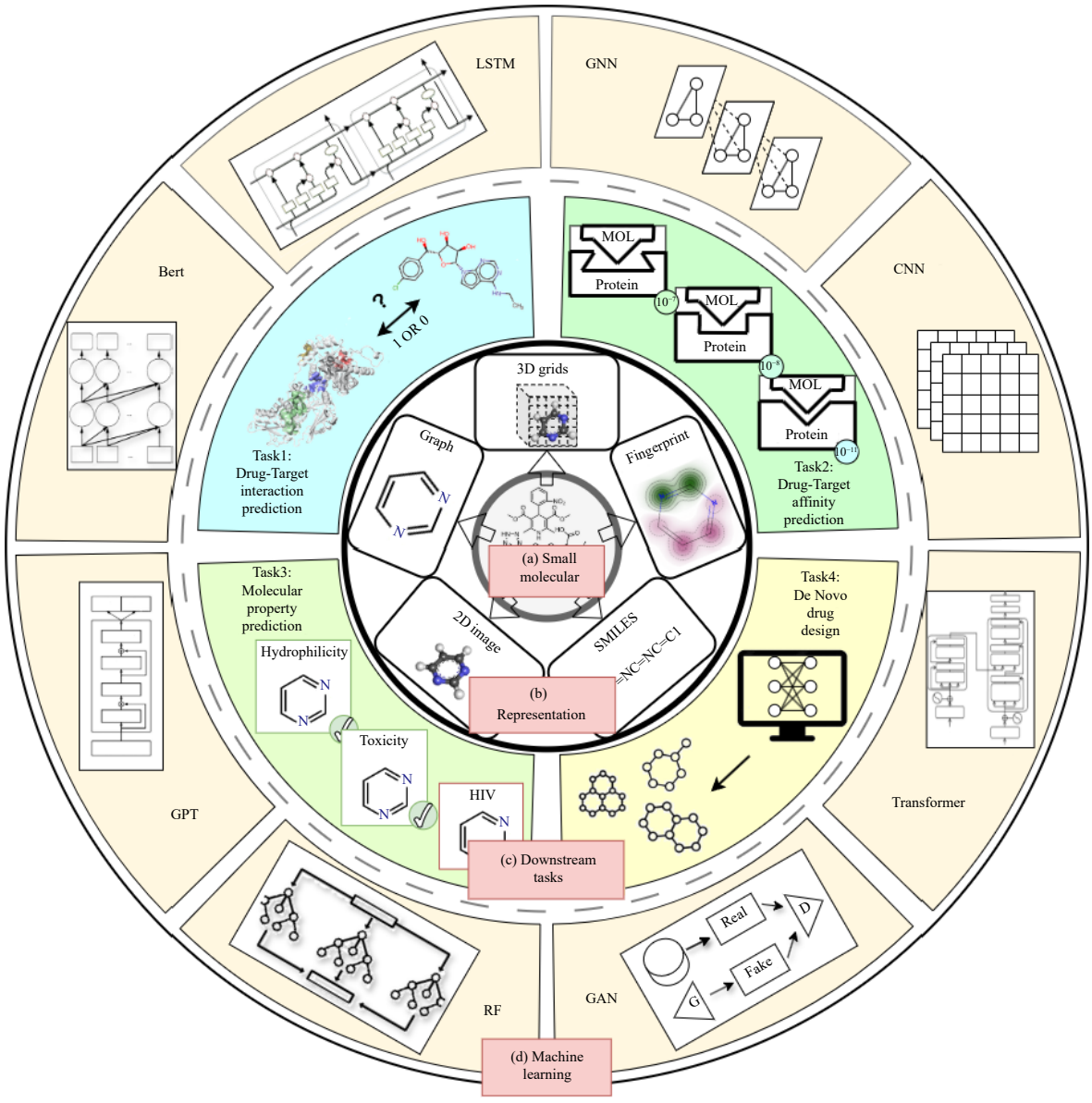

Fig. 1 The overall flowchart from the representation of small molecules to their relevant downstream tasks. (a) Small molecules. (b) Methods of small molecule representation. (c) Four downstream tasks. (d) The machine learning methods usually applied in the downstream tasks.

and cross-database links to resources like ChEMBL and DrugBank. However, it lacks comprehensive clinical trial data (e.g., patient-level outcomes) and underrepresents emerging areas like gene therapies and rare disease drugs 16. While open access, advanced features may require institutional subscriptions, and real-time updates are absent, limiting its utility for time-sensitive applications [32].

2.4. PubChem 2023

The PubChem 2023 represents a widely utilized chemical information dataset in drug development research. The dataset primarily contains chemical structures, identifiers, chemical and physical properties, biological activity, patents, health, safety, and toxicity data of small molecules and larger molecules. Additionally, it incorporates chemical entities such as siRNA, miRNA, lipids, carbohydrates, and chemically modified biopolymers. Certain compounds lack clear target annotations. The database includes bioassay results from high-throughput screening, potentially affected by varying experimental conditions [7].

2.5. DrugCentral 2023

DrugCentral 2023 is an open-access digital repository that integrates diverse drug-related data. Extensively used in pharmaceutical and biomedical research, it contains information on chemical structures, molecular physicochemical descriptors, patent status, biological activities, and molecular targets. For some drugs, target annotations are limited to target names without corresponding binding affinity data, which may constrain certain analytical applications [33].

2.6. TTD

The TTD is a target dataset, commonly used for drug target and drug discovery research. It features molecular interactions, human system characteristics, and cell-based expression variation classification. It primarily includes data on drugs and their related targets, clinical trial targets, patent targets, and literature-reported targets with established drug-like characteristics. Its strengths include detailed target validation status (e.g., clinical efficacy data), disease association networks, and cross-referencing to omics resources. However, it underrepresents emerging targets in precision medicine (e.g., cell therapies) and lacks real-time updates for rapidly evolving clinical trials [34].

2.7. ZINC-22

The ZINC22 is a comprehensive database of commercially available small molecules, providing standardized 3D structures, physicochemical properties (e.g., logP, molecular weight), and vendor information for over 2.3 billion compounds. It includes specialized subsets like FDA-approved drugs and fragment libraries, optimized for virtual screening and AI-driven drug discovery [35]. Its strengths include diverse chemical space coverage, machine-readable SMILES/3D formats, and direct links to vendor purchasing info. However, it lacks curated biological activity data, may contain structural duplicates, and some listed compounds are out-of-stock, limiting functional validation and real-world applicability.

2.8. PROMISCUOUS 2.0

PROMISCUOUS 2.0 serves as a comprehensive dataset for drug repositioning research, facilitating the identification of novel indications for existing compounds. The dataset encompasses information on drugs, targets, and their interactions, as well as relationships between drugs and side effects. However, certain drug-target-adverse reaction associations within the database are derived from text mining methodologies and lack experimental validation [36].

2.9. ChEMBL in 2017

ChEMBL in 2017 represents an extensive bioactivity dataset essential for drug discovery research. The database incorporates bioactive compounds with their two-dimensional structures, computed properties, and abstracted bioactivity data. Furthermore, it encompasses information pertaining to ontology annotations, target detection, and targets [37].

2.10. SuperDRUG 2

SuperDRUG 2 comprises a comprehensive collection of approved and marketed drugs, serving as a valuable resource for drug development. The dataset encompasses detailed information on drugs, including their chemical structures, regulatory information, indications, drug targets, side effects, physicochemical properties, pharmacokinetics, and drug-drug interactions [38].

2.11. SIDER 4

SIDER 4 functions as a repository of drugs and their associated side effects. Adverse drug reactions documented during clinical trials provide essential human phenotype data. The dataset, frequently utilized in new drug development and adverse drug reaction research, contains comprehensive information on drugs, side effects, and their interactions. Its notable features include standardized MedDRA coding, drug-ADR association confidence scores, and cross-links to pharmacovigilance databases. However, limitations include delayed real-time ADR updates (last major update in 2015), absence of patient-level demographic data, and insufficient representation of ADRs from post-marketing surveillance or non-Western populations, constraining its utility for contemporary safety monitoring [39].

2.12. KEGG-DRUG

KEGG represents an extensive dataset of genes and genomes, incorporating biological interpretations of large-scale molecular datasets. Within this framework, KEGG-DRUG serves as a specialized dataset for drug development research. It contains an extensive collection of approved drugs, including information on drug targets and drug metabolism. The metabolic data is categorized by metabolic enzymes and transporters, along with their respective substrates, inhibitors, and inducers [40].

2.13. STITCH 5

STITCH 5 functions as a dataset focused on protein-small molecule interactions, widely utilized in chemical interaction studies. The database primarily contains information on proteins, small molecules, and their interactions, including binding affinity data [41]. This resource integrates established and predicted chemical-protein interaction data from multiple sources, providing a comprehensive platform for molecular association studies. The system enables network visualization of interactions with binding affinities and allows filtering based on tissue expression patterns, enhancing context-specific interaction analysis. Nevertheless, predicted interaction accuracy varies, and the integration of diverse data sources may introduce inconsistencies, requiring careful interpretation. Moreover, effective utilization of its complex outputs typically requires substantial biological knowledge.

2.14. BindingDB in 2015

As of 2015, BindingDB served as a publicly accessible repository specializing in experimentally determined protein–small molecule interaction data. The database contained over one million records, derived primarily from peer-reviewed publications and, to an increasing extent, from U.S. patents. These records encompassed interactions between nearly 500 000 small molecules and thousands of proteins, with binding affinities measured using diverse experimental techniques, including enzyme inhibition and kinetic assays, isothermal titration calorimetry, NMR spectroscopy, and radioligand competition assays. Each entry integrated textual annotations, chemical structures, protein sequences, and quantitative affinity data linking proteins to small molecules. However, BindingDB has notable limitations. The absence of updates beyond 2015 results in the exclusion of recently identified target–ligand interactions and newly developed assay technologies. In addition, its limited inclusion of non–small molecule binders and inconsistencies in data validation standards may affect its reliability and applicability for current research needs [42].

2.15. TCMSP

TCMSP functions as a dataset of TCMs, extensively utilized in research and development related to TCM. The database encompasses information on drugs and their chemical properties, absorption, distribution, metabolism, and excretion (ADME) characteristics, drug similarity, drug targets, associated diseases, and interaction data. However, the database suffers from infrequent updates, resulting in the absence of current data [43].

2.16. CancerDR

CancerDR is a publicly accessible dataset frequently utilized in cancer and drug target mutation research. The database encompasses information on 148 anticancer drugs and their pharmacological profiles across 952 cancer cell lines. For each drug target, it provides detailed information, including natural variant sequences, mutations, tertiary structures, and alignment profiles of mutants/variants. Furthermore, CancerDR incorporates several web-based tools. The database facilitates the identification of genetic alterations in drug target genes and residues associated with drug resistance, while also enabling users to identify versatile drug molecules effective against various cancer cells. Nevertheless, the database has remained unupdated for a considerable period, and its mutation data remains incomplete [44].

2.17. SuperTarget

SuperTarget is a dataset of drugs and targets, frequently employed in studying drug-target-pathway relationships for novel drug development. The database primarily contains information on drugs and their indications, adverse reactions, metabolism, pathways, and gene ontology terms associated with target proteins. It combines experimental and predicted DTI data, providing a comprehensive resource for therapeutic target identification. However, certain enzyme interactions within the database have not incorporated recent clinical research findings [45].

Several datasets exclusively contain small molecule data. The structural and physical properties of these small molecules can be analyzed to extract valuable features. The correlation between molecular structure and physical properties enables the training of deep neural networks for property prediction. In drug development, these features can function as templates for generating new small molecules with comparable characteristics in MG tasks [46]. Other datasets incorporate both small molecule data and information about interacting entities. This relational data supports tasks such as DTI or DTA prediction, facilitating drug screening in pharmaceutical development [47]. Table 1 presents a concise overview of these datasets, including their descriptions, associated downstream tasks, update timelines, and accessibility links. The downstream tasks encompass DTI prediction, DTA prediction, DP prediction, and MG.

3. Representations of small molecules

The representations of small molecules can be categorized into two main groups: basic digital representations and learned representations. Basic digital representations convert molecular structures into machine-readable formats, encompassing string format, graph format, and 2D image/3D grid format. Learned representations, designed for AI-based downstream tasks, incorporate comprehensive and complete feature information. These learned representations include string-based, graph-based, 2D image/3D grid-based, hybrid, and pre-trained large model representations. Fig. 2 illustrates the overall representations of small molecules.

3.1. Molecular representations

3.1.1. String representation

(1) IUPAC

The systematic nomenclature developed by the IUPAC represents the earliest standardized approach for molecular description. This hierarchical naming system employs prefixes, suffixes, and numerical locants to unambiguously characterize chemical structures with atomic precision [16]. The nomenclature maintains chemical intuitiveness, enabling direct human interpretation of structural features, including functional groups and substitution patterns. For example, the IUPAC nomenclature of isoflavone is 3-Phenyl-4H-chromen-4-one. Nicotine, a naturally occurring alkaloid, has the International Union of Pure and Applied Chemistry (IUPAC) name (S)-3-(1- Methylpyrrolidin-2-yl)pyridine.

This conventional chemical nomenclature, however, presents several limitations. The intricate rule system requires complex stereochemical descriptors (e.g., R/S, E/Z configurations), often generating verbose representations for complex molecular architectures. Additionally, the methodology demonstrates incompatibility with modern computational workflows, requiring preprocessing through specialized cheminformatics tools (e.g., Open Babel) to convert nomenclature into machine-readable formats suitable for algorithmic processing.

(2) SMILES

The SMILES encodes the structures of small molecules as linear strings composed of ASCII characters. Unlike the hierarchical and systematic conventions of IUPAC nomenclature, SMILES provides a compact, machine-readable representation that captures molecular connectivity in a sequential format. It represents molecular structure through strings that specify atom types and their connections. Atoms are denoted by their atomic symbols, while chemical bonds are represented by punctuation marks: single bonds as "-", double bonds as "=", and triple bonds as "#". Ring structures are indicated by numbers following the starting and ending atoms, while branches are denoted by parentheses, as illustrated in Fig. 2(a). This representation method efficiently captures atomic composition and molecular connectivity information, making it a prevalent input format for chemical small molecules. For example, isoflavone's SMILES notation is O=C1C=COC2=CC=CC=C12C1=CC=CC=C1, while nicotine's is CN1CCC[C@H]1c2cccnc2. A significant advantage of SMILES is its canonicalization capability, where methods like Canonical SMILES generate unique strings for each molecular structure. This normal-

Table 1 The representative databases related to small molecules

Dataset	Data description	Downstream Tasks	Update
DGIdb 5.0 [31]	~ 70 000 drugs and their related gene information	DTI, DP, MG	2024
DrugBank 6.0 [6]	22 685 drugs involving 1 413 413 drug-drug interactions	DTI, DP, MG	2024
IUPHAR/BPS in 2024 [32]	3 039 protein targets, 12 163 ligand molecules	DTI, DP, MG	2024
PubChem 2023 [7]	111 892 547 compounds, 103 988 genes, 185 153 proteins	DTI, DP, MG	2023
DrugCentral 2023 [33]	3 952 small-molecule data, 704 targets, and 1 640 interaction relationship data	DTI, DTA, DP, MG	2023
TTD [34]	114 clinical trials, 212 preclinical trials, 1 479 literature-reported druggable targets, and 39 862 drugs	DP, MG	2023
ZINC-22 [35]	~ 37 000 000 000 compounds	DTI, DP, MG	2023
PROMISCUOUS 2.0 [36]	~ 1 000 000 small molecules, 110 000 side effects, 3 000 000 drug-target interactions	DTI, DTA, DP, MG	2021
ChEMBL in 2017 [37]	1 665 198 compound structures (~ 11 000 targets, 9 052 proteins)	DTI, DP, MG	2017
SuperDRUG 2 [38]	4 587 types of active pharmaceutical ingredients	DP, MG	2017
SIDER 4 [39]	1 430 drugs and 14 064 drug-adverse reaction pair data.	DP, MG	2016
KEGG-DRUG [40]	12 320 drugs	DP, MG	2016
STITCH 5 [41]	~ 9 600 000 proteins and ~ 430 000 compounds	DT, DP, MG	2015
BindingDB in 2015 [42]	~ 490 000 small molecules with ~ 1 100 000 affinity data	DTI, DTA, DP, MG	2015
TCMSP [43]	499 types of medicine, 29 384 components, 3 311 targets on 837 related diseases	DP, MG	2014
CancerDR [44]	148 types of anticancer drugs in 952 cancer cell lines	DP, MG	2013
SuperTarget [45]	~ 6 000 target proteins, ~ 330 000 interactions with about 196 000 compounds	DTI, DP, MG	2011

ized form employs an atom-prioritizing algorithm, establishing a one-to-one correspondence between molecular structure and representation [48]. The format's space efficiency enables faster deep neural network training with reduced memory requirements. However, SMILES representation shows limited sensitivity to structural similarities between molecules, where minor structural variations can produce significantly different strings, potentially compromising feature learning in deep neural networks.

(3) SMILES modification

Several extended molecular representation methods have been developed to address the limitations of SMILES notation.

● The SMARTS

SMARTS extends the original SMILES by incorporating wildcard atoms (*), logical operators (&, !), and property modifiers (e.g., charge states, ring sizes), enabling sophisticated substructure matching for applications such as pharmacophore identification - albeit at the cost of increased syntactic complexity.

● The SMILES Reaction Specification (SMIRKS)

SMIRKS extends this framework through the introduction of reaction arrows (>>) and atom mapping indices (: 1,: 2), permit-

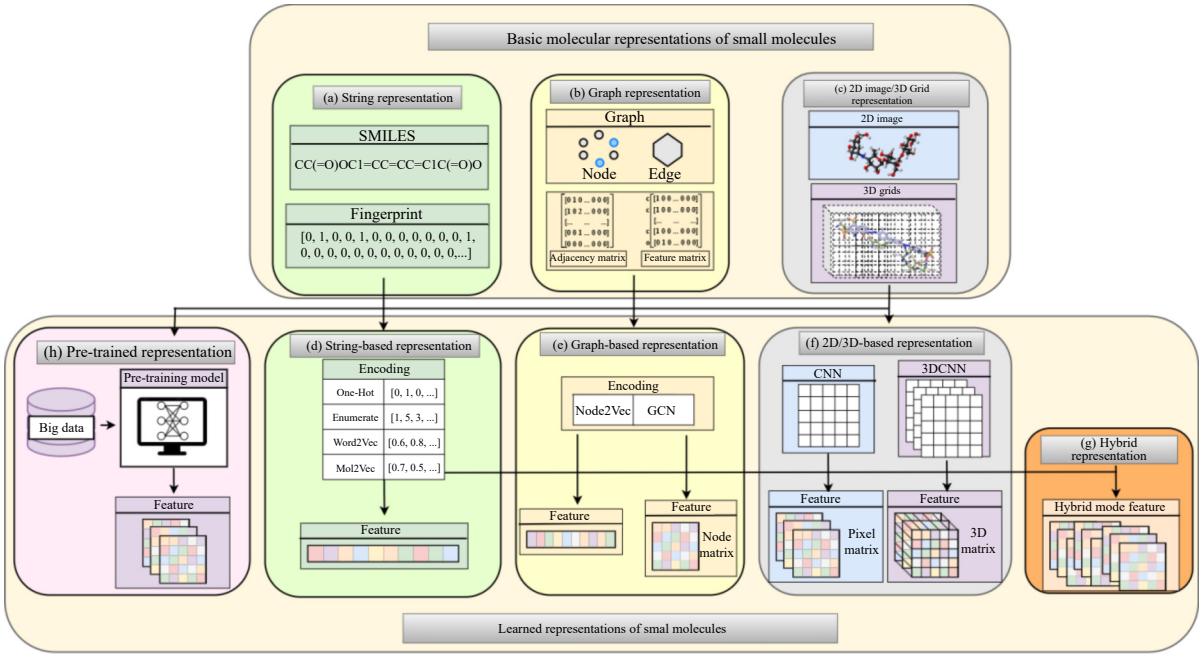

Fig. 2 The basic digital representations and learned representations of small molecules. (a) String representation of small molecules; (b) Graph representation of small molecules; (c) 2D Image/3D Grid representation of small molecules; (d) String-based representation; (e) Graph-based representation; (f) 2D image/3D grid representation; (g) Multi-modal hybrid representation of small molecules; (h) Pre-trained representation.

ting precise representation of chemical transformations, though its utility diminishes with increasingly complex reaction systems.

● OpenSMILES

OpenSMILES establishes standardized syntactic conventions that address ambiguities in the original specification, particularly regarding hydrogen atom representation, chirality descriptors, and aromaticity definitions [49].

● SELFIES

While SMILES syntax is susceptible to invalid strings through random character modifications, SELFIES ensures grammatical validity by separating the storage of branch and ring information, facilitating the generation of chemically viable molecules despite random mutations. The SELFIES string for isoflavone is: [C][=C][O][C][=C][Branch1] [Ring1][=C][Branch1][Ring2][C][=C][C][=C][C][=C][Ring1][C][=O][Ring2]. The SELFIES encoding for nicotine is: [N][C][Branch1][Ring1][C][C][C][C][=C][Ring1][C][=C][C][=C][N][=C]. Additionally, SELFIES exhibits superior validity rates and diversity in generated molecules. In the DRAGONFLY [50], for *de novo* drug design tasks targeting proteins such as PPARγ, LXRβ, RARα, BRAF, BTK, and JAK2, the SELFIES-based model demonstrated better performance than the SMILES-based model regarding the uniqueness and novelty of generated molecules.

(4) Molecular Fingerprint

Molecular fingerprinting encodes molecular structures into bit vectors, with various fingerprint representations based on different encoding rules, analogous to human fingerprints. Dictionary-based molecular fingerprints utilize binary positions, where 1 indicates the presence of a predefined functional group, and 0 denotes its absence. In isoflavone's molecular fingerprint, structural fragments such as the keto group, benzene ring, and pyran ring correspond to specific bit positions. For nicotine's molecular fingerprint, structural motifs like the pyridine ring and methyl group map to designated bit positions. Path-based molecular fingerprints generate patterns by hashing paths from atoms to specified positions. Extended Connectivity Fingerprints assign integer identifiers to heavy atoms, merging and updating atoms within specified radii to create feature lists. Pharmacophore molecular fingerprints describe molecular interactions, incorporating chemical information of atom types and structural data such as shortest path tuples between atoms. While molecular fingerprints serve as input data for deep neural network training, their limited data quality and lack of standardization restrict their effectiveness [51].

3.1.2. Graph representation

The molecular graph representation is a widely adopted method for encoding the structural and chemical features of small molecules. The representation typically employs atoms as vertices and inter-atomic bonds as edges to form graph structural information. This format represents atomic types and structural information in a graph data structure. In isoflavone's molecular graph, carbon (C) and oxygen (O) atoms function as nodes, with edges representing single, double, and aromatic bonds between atoms. Similarly, in nicotine's molecular graph, C and nitrogen (N) atoms form nodes, with edges indicating single, double, and aromatic bonds. This representation method has evolved significantly alongside deep neural networks that process graph structure information, such as GNNs and graph convolutional neural networks (GCNs), as illustrated in Fig. 2(b).

Molecular graph representation offers superior capture of structural features compared to molecular sequences, retaining more molecular information. However, it presents challenges in handling cyclic structures during node-edge data conversion. One viable approach involves converting molecular graphs into tree graphs, transforming cyclic graphs into acyclic forms [52]. For computational representation, the graph requires abstraction into a set of nodes V and edges E. Typically, matrices represent node feature vectors and edge feature vectors, while an adjacency matrix indicates node connectivity and directionality. The node set encompasses atomic type information and 3D data, including atomic coordinates and bond angles.

3.1.3. 2D image/3D grid representation

The advancement of computer vision and enhanced deep neural network capabilities in processing 2D images has influenced molecular representation, leading to attempts at training molecules represented as 2D images. These high-quality 2D images can contain atomic types and molecular structure information, which deep neural networks like CNNs can extract, as demonstrated in Fig. 2(c). While high-quality 2D images can contain substantial information, limitations include the scarcity of high-quality 2D image datasets and the dependence on network performance for effective feature extraction.

The aforementioned methods rarely incorporate three-dimensional positional information, and utilizing this information effectively remains challenging even in 3D molecular maps. While placing atoms in a 3D grid representation yields favorable results in 3D CNN applications, this approach presents certain limitations, including increased computational resource requirements and network sparsity.

Beyond conventional representations, specialized molecular file formats can directly encode three-dimensional structural information.

● Structure-Data File

The SDF format utilizes text-based records where each molecular entry explicitly enumerates atomic coordinates and bonding connectivity while accommodating auxiliary property data.

● MOL2

The MOL2 format extends this capability by incorporating advanced molecular features - including partial atomic charges, atom typing parameters, and force field specifications - while maintaining clear substructure differentiation.

3.2. Learned representation

3.2.1. String-based representation

Molecular representations in string formats require transformation of their symbolic notations into vector spaces suitable for neural network computation. As shown in Fig. 2(d), common conversion methods include enumeration, one-hot encoding, word2vec, and mol2vec, specifically utilized in biological applications. One-hot encoding represents a fundamental computational encoding method. This approach initially segments the molecular sequence into basic units, typically letters or symbols. A comprehensive vocabulary of these units is constructed and indexed. Subsequently, each unit receives a vector representation with length equal to the total unit count, where the corresponding unit position contains 1 while all others contain 0. Similar approaches include enumeration, which directly assigns numerical values to letters and symbols after vocabulary construction. While one-hot encoding provides a straightforward method for vector representation of molecular sequences, it presents limitations in storage efficiency due to sparse vectors containing numerous zero values and the inability to capture unit similarity relationships.

To address storage efficiency and enhance feature vector information content, some researchers employ the word2vec method from NLP to convert molecular sequences into feature vectors [53]. The word2vec method operates by predicting central vocabulary using contextual information and vice versa.

Word2vec-transformed vectors can contain real numbers within specified ranges rather than solely integers. This signific-

antly reduces vector dimensionality and storage requirements. Additionally, similar word2vec-transformed feature vectors maintain proximity in feature space, effectively representing relationships between different feature vectors. Fig. 3 illustrates a basic word2vec workflow. Mol2vec, widely employed for converting molecular sequences into feature vectors, adapts this process by substituting natural language training data with atomic data.

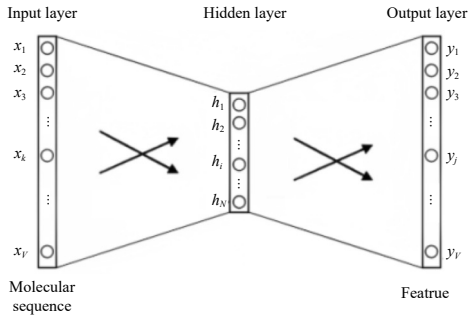

Fig. 3 A simple word2vec workflow.

(i) LSTM in RNNs represents a neural network architecture capable of processing sequential information with bidirectional relationships. Its information transmission units selectively retain or discard preceding sequence information. For instance, Zhang et al. proposed a generative network combining LSTM and dense full CNNs [54]. Gupta et al. proposed a network utilizing LSTM for recursive drug generation [55]. Wu et al. proposed a neural network incorporating bidirectional LSTM and a multi-step attention mechanism [56]. Wang et al. developed a network integrating position-specific scoring matrix and drug molecule substructures [57], where small molecules are encoded into 881-dimensional feature fingerprints based on dictionary input for LSTM training. These approaches employ LSTM training after obtaining appropriately formatted SMILES representation data.

(ii) For SMILES-represented molecular data, Transformer enhances embedding with contextual substructure information [58]. For example, Huang et al. proposed a molecular interaction Transformer network utilizing molecular substructures [59]. After molecular substructure division, a learnable substructure lookup dictionary is initialized, followed by direct embedding enhancement through the Transformer.

(iii) SMILES-represented data input to GPT generates rich text descriptions through optimized prompts. For instance, Balaji et al. proposed a GPT-based text description molecular property prediction network [60]. To obtain molecular description data, SMILES representations are input into GPT to generate corresponding rich text descriptions for training.

(iv) SMILES-represented data combines with CNNs for feature extraction. For example, Kalemati et al.proposed a network based on alignment similarity measures [61]. The molecular structures underwent length standardization before CNN training input.

(v) Generative adversarial networks (GANs) can be trained on real small-molecule SMILES representation data [62] to generate novel molecular structures. For example, Zhao et al. developed a semi-supervised GAN [63] that employs SMILES representations of small molecules as input data, enabling the model to generate predicted SMILES strings corresponding to potential new compounds.

(vi) Early ML algorithms, combined with string representations containing structural and functional group information, demonstrate notable effectiveness. For instance, Chen et al. proposed an SVM and RF network based on the molecular association networks [64], achieving favorable results through the construction of Morgan fingerprints in both SVM and RF implementations.

3.2.2. Graph-based representation

In graph-structured data analysis, node feature vectors are initially obtained through various embedding methods, including Node2Vec or GCNs. Node2Vec converts each node's local topological structure into sequence information using Biased Random Walks, with hyperparameters favoring either breadth-first or depth-first search. The sequence undergoes truncation at a specific length, followed by embedding representation learning based on Word2Vec. Molecular graph representation typically employs graph-centric neural networks such as GNNs and GCNs. In GCN implementations, graph structure data are stored in three matrices: the node matrix, edge matrix, and adjacency matrix. The update process involves aggregating neighboring node features, incorporating individual features, and implementing cyclic updates. During neighbor node feature aggregation, the adjacency matrix represents connections between nodes with 1 and the absence of connections with 0. Various GNN models, including GAT, implement different aggregation and update mechanisms, incorporating weights and attention mechanisms to optimize graph structure information capture and node feature extraction. This approach accommodates molecular 3D structural information, where some methods utilize atomic 3D coordinates to form additional matrices complementing the node, edge, and adjacency matrices. The evolution of GNNs and similar neural networks has led to increased adoption of molecular graph representations for small molecules, combined with diverse ML methodologies.

(i) While string representations struggle to reflect structural similarities between different small molecules, molecular graphs can be generated from SMILES representations and trained using GNNs [65]. Jiang et al. introduced a drug target affinity prediction network utilizing GNNs [66]. Zhai et al. developed a network combining dynamic graph attention with bidirectional LSTM [67]. Nguyen et al. proposed networks based on GCNs, GAT, GIN, and GAT-GCN [68-71]. Qi et al. integrated a graph learning module with soft adjacency matrix refinement for protein and drug molecule graphs [72]. Bai et al. presented a domain-adaptive bilinear attention network [73]. These approaches utilize SMILES input for small molecules, converting them into molecular graphs of atoms and edges for training, considering their natural properties.

(ii) Enhanced structural information integration in GNNs enables additional neural network insights. For example, Fang et al. developed a geometrically enhanced molecular representation learning neural network [74], utilizing dual graphs: one treating atoms as nodes and bonds as edges, and another treating bonds as nodes and bond angles as edges.

(iii) The integration of Bert's masking concept into GNNs enhances model generative capabilities. Xia et al. developed a deep neural network for atomic representation pre-training based on masking and VQ-VAE [75], implementing molecular graph masking and contrastive learning between different mask ratios to improve performance.

3.2.3. 2D image/3D grid-based representation

Advances in small molecule data extraction have facilitated the utilization of high-quality 2D image and 3D grid data. These representation methods effectively capture spatial structural information and readily convert into computer-training-compatible formats.

(i) In image-based molecular representation, CNNs are extensively employed as primary feature extractors in computer vision tasks, including the analysis of 2D depictions of small molecules. Rifaioglu et al. designed a CNN architecture based on a two-dimensional structural composite representation [76], training the network directly on small-molecule 2D images. Zeng et al. [77] developed a self-supervised pretraining framework in which an encoder processes both local and global structural features ex-

tracted from small-molecule 2D images.

(ii) For 3D molecular representation, Casey et al. employed multi-resolution molecular 3D grids as multi-channel input for CNN-based modeling [78]. To address the inherent sparsity of such grids, Kuzminykh et al. introduced a smoothing algorithm to interpolate spatial regions between atoms, thereby enhancing feature density and improving CNN training performance [79].

3.2.4. Multi-modal-based representation

Each small molecule representation method exhibits distinct limitations. Researchers have explored combining multiple molecular representations. This multi-modal fusion approach, termed "Hybrid Mode" in this study, integrates the advantages of various representation methods. Wang et al. [80] proposed a multi-level message-passing GNN that inputs both molecular graph and SMILES representations, optimizing generation performance through a masking method. Clevert et al. [81] developed a network converting molecular descriptions into SMILES by combining 2D image and SMILES representations to infer molecular structure. Li et al. [82] created a neural network for fine-tuning pre-trained Bert using molecular fingerprint and SMILES representations, predicting SMILES substructures through masked training. Stärk et al. [83] designed a network predicting molecular geometry through 2D structure by processing 2D molecular graphs and 3D molecular point clouds through GNNs and 3DGNNs, respectively, before merging. Xia et al. [84] proposed a multi-modal network integrating knowledge graphs, gene expression maps, and gene expression features for training. Liu et al. [85] developed a model predicting and optimizing molecules using natural language processing, taking molecular graphs and text descriptions as input. Notably, MoleculeSTM employs contrastive learning to align molecular structures with textual descriptions in a joint embedding space, integrating chemical structure with textual knowledge. This enables text-based instruction tasks without fine-tuning. Edwards et al. [86] proposed a cross-modal molecular natural language retrieval network inputting Mol2vec-converted molecular graph sequences and molecular text descriptions into a transformer. Qi et al. [87] developed an unsupervised learning network for drug target binding regions, merging molecular graph and extended connectivity fingerprint representations through an attention mechanism. Lu et al. [88] demonstrated enhanced predictive performance by integrating diverse molecular representations through MMFDL, using Transformer-Encoder for SMILES vectors, BiGRU for ECFP fingerprints, and GCNs for molecular graphs. Xie et al. [89] combined MACCS keys and ECFP fingerprints to construct comprehensive molecular representations, significantly improving predictive performance.

3.2.5. Pre-trained LLM-based representation

LLMs have emerged as a transformative technology in drug discovery and chemoinformatics [90]. This approach utilizes deep learning frameworks to automatically extract features and perform predictive tasks from molecular data. The methodology provides two key advantages: it eliminates manual feature engineering through direct learning of latent representations, and enhances feature representation generalizability through pre-training on large unlabeled datasets followed by downstream task fine-tuning. However, limitations include substantial computational requirements for initial model training, necessitating GPU clusters and extensive datasets that exceed traditional approach requirements. For instance, Mol-LlaMA [91] integrates a 2D-3D molecular encoder combining molecular structure and language instruction information, enhancing molecular understanding and property prediction effectiveness. Mol-LLM [92] incorporates two GNN architectures (GINE and TokenGT), enhancing molecular representation capabilities through functional group prediction and SELFIES reconstruction pre-training tasks. Additional recent pre-trained large models for molecular applications are presented in Table 2.

4. Downstream tasks related to small molecule representation

Computational methods in drug discovery have significantly reduced development time and costs. These methods focus on binary DTI prediction, regression-based DTA prediction, drug pharmacokinetic property forecasting, and *de novo* small MG. The following sections examine these AI-driven drug research tasks in detail.

4.1. DTI prediction

DTI analysis represents a fundamental aspect of pharmaceutical research. The primary objective is to determine the presence of interactions between drugs and their targets. These experimental findings help minimize unintended interactions between drugs and off-target proteins during applications, thereby reducing adverse effects. To enhance efficiency and reduce costs, com-

Table 2 List of popular small molecule pre-trained Large Language Models (LLMs) in recent years.

Name	Training Dataset	# Parameters	Based model	Time
ChemBERTa-2 [93]	PubChem	46M	RoBERTa	2022
MFBERT [94]	ZINC, PubChem	88M	RoBERTa	2022
Molformer [95]	ZINC, PubChem	110M	Transformer	2022
Chemformer [96]	ZINC	230M	Transformer	2022
X-MOL [97]	ZINC	110M	Transformer	2022
Uni-Mol [98]	ZINC, ChemBL, PDBbind	48M	Transformer	2023
RetroSynth-Diversity [99]	USPTO, Pistachio	12M	Transformer	2023
SELFormer [100]	ChemBL	87M	RoBERTa	2023
iupacGPT [101]	PubChem	1500M	GPT	2023
MolecularGPT [102]	ChEMBL, QM9	\	LLaMA	2024
MolTC [102]	ZINC	12M	Galactica	2024
Mol-LLaMA [91]	PubChem, Mol-Instructions, ChEBI-20	\	Llama	2025
Mol-LLM [92]	MoleculeNet, ChEBI-20, PubChem, USPTO	\	Mistral	2025

putational and AI technologies are increasingly utilized in this field. Deep neural networks can effectively perform this binary classification task of interaction prediction. The training process typically employs verified DTI data, operating under the principle that structurally similar drugs tend to interact with structurally similar targets [103].

Fig. 4 illustrates the standard workflow for DTI. This process typically begins with datasets containing interaction or affinity information. Proteins are converted into feature vectors, while small molecules are represented through various previously discussed methods, yielding feature vectors or matrices. When utilizing separate protein and small molecule data, interaction or affinity information is typically incorporated after obtaining feature vectors through pre-trained models. Some approaches first combine the feature representations of proteins and small molecules before processing them through neural networks such as LSTM, CNNs, or Transformer for training to extract hidden information. Alternative approaches process protein and small molecule feature representations separately through neural networks to obtain their respective hidden information before combination. These hidden representations are then processed through additional ML neural networks, such as fully connected networks, to generate binary classification predictions. The effectiveness is typically evaluated through binary classification metrics. Several commonly used methods are illustrated in Fig. 5.

4.1.1. Basic ML methods for DTI prediction

Early DTI tasks implemented fundamental ML methods. Manually engineered features demonstrate effectiveness in small-scale datasets. Chen et al. developed support vector machine and RM models that achieved high predictive accuracy by integrating molecular attribute features with behavioral characteristics. Additionally, An et al. proposed a heterogeneous network random walk approach combined with gradient-boosted decision trees that demonstrated exceptional prediction performance on benchmark datasets [104].

4.1.2. CNN and RNN-based models for DTI prediction

Following their successful application in various domains, CNN and RNN architectures have matured and demonstrate effective classification capabilities when coupled with fully connected networks. Chen et al. developed a molecular association network that enhances predictive performance through multi-dimensional feature integration. An et al. introduced a heterogeneous network random walk approach offering novel perspectives for complex target prediction. Ren et al. constructed a multimodal deep neural network that effectively interprets semantic information from protein-drug-disease networks. Rifaioglu et al. developed a two-dimensional structural representation model demonstrating exceptional performance in predicting interactions with critical targets, including MAPK14 and JAK1. Wang et al. developed an LSTM network combining position-specific scoring matrices with drug molecular substructure fingerprints, while Yu et al. introduced a Bi-LSTM and attention-based heterogeneous GNN, both achieving superior predictive capabilities [105].

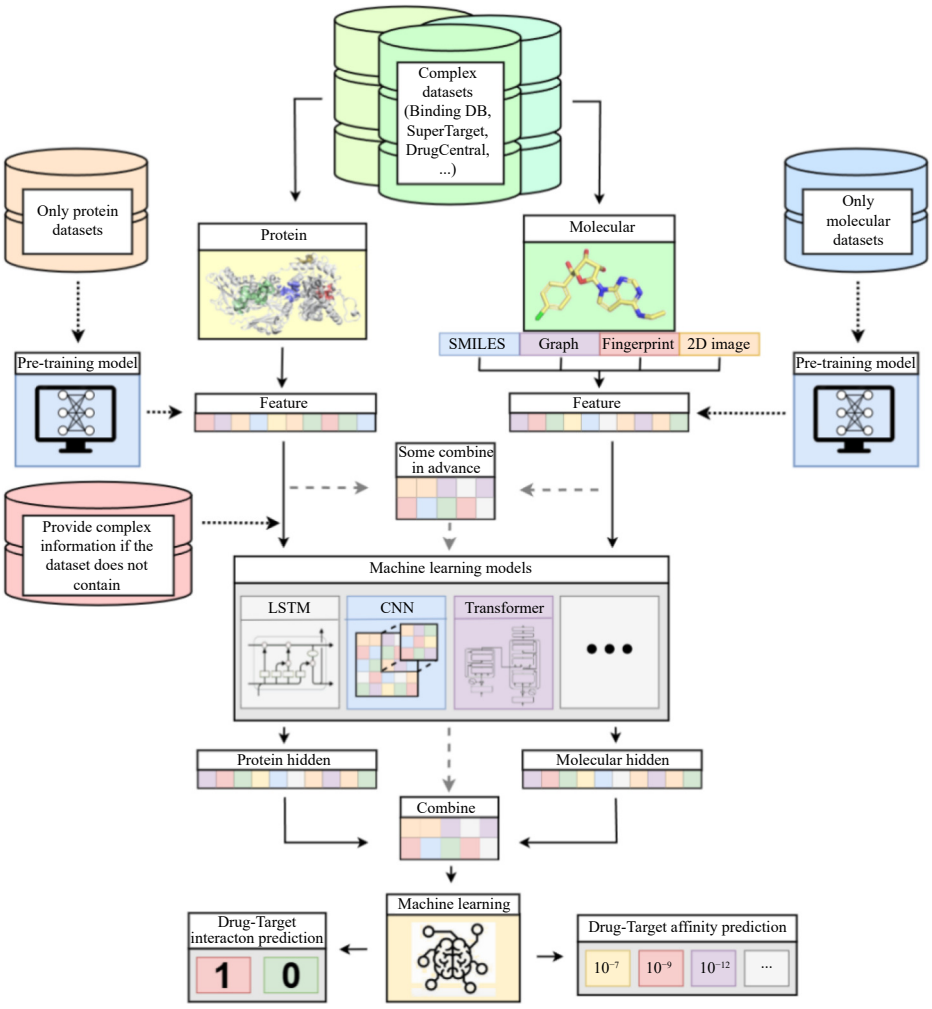

Fig. 4 The common process adopted for DTI and DTA prediction.

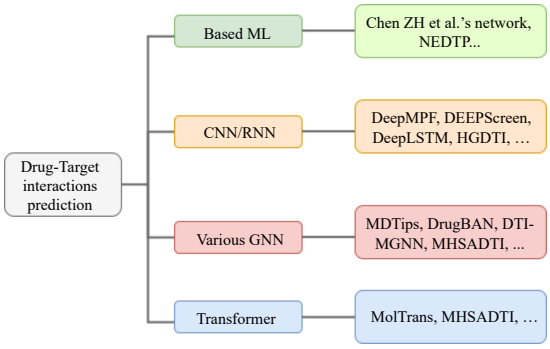

Fig. 5 Common DTI prediction models.

4.1.3. Various GNN-based models for DTI prediction

Current research trends focus on structuring data into graphs for training with various GNNs. Xia et al. developed MDTips, a knowledge graph-integrated neural network that effectively combines structural data with gene expression maps and biological knowledge graphs, demonstrating robust predictive performance. Bai et al. introduced DrugBAN, a domain-adaptive bilinear attention network achieving cross-dataset generalization through dynamic feature interaction weighting. Li et al. advanced graph-based approaches with DTI-MGNN, a multi-channel GNN processing both topological and feature graphs through dedicated GAT and GCN pathways [106]. Cheng et al. proposed an end-to-end neural network utilizing graph attention networks and multi-head self-attention mechanisms for processing small molecules and proteins. This method demonstrated exceptional performance across human, C. elegans, DUD-E, and DrugBank datasets, highlighting its effectiveness in prediction tasks [107].

4.1.4. Transformer-based models for DTI prediction

The transformer, a significant precursor to GPT and Bert, finds increasingly widespread application. Huang et al. proposed a molecular interaction transformer network utilizing molecular substructures for modeling. This network demonstrated exceptional performance in DTI tasks across BioSNAP, DAVIS, and BindingDB datasets, illustrating the model's adaptability across diverse data environments. Additionally, Cheng et al.'s neural network incorporates Transformer components, further demonstrating this architecture's potential in processing complex biomedical datasets.

4.2. DTA prediction

DTA represents a significant task in drug research, investigating the binding affinity relationship between drugs and targets. Unlike the DTI task, which employs binary classification to determine DTI, DTA provides continuous values indicating the strength of affinity between specific drugs and targets, offering more valuable insights for drug development [108]. Networks trained for DTA tasks demonstrate practical applications in drug development. For example, DTIAM [109] successfully identified approved EGFR inhibitors, accurately predicted four approved CDK4/6 inhibitors, and its predicted compounds, including Imatinib mesylate and Alvocidib, received validation through external databases. Furthermore, specialized methodologies have emerged for predicting herb-target interactions in TCM. HTINet [110] exemplifies this through its approach of learning low-dimensional feature representations of herb and protein nodes, implementing various ML models (including KNN, SVM, LR, DT, RF, and GBDT) to predict herb-target interactions with notable effectiveness. HTINet2 [111] further advances the prediction capabilities of herb-target interactions by incorporating TCM knowledge graphs, residual GNNs, and supervised learning, establishing a robust framework for modernizing TCM research. Fig. 4 illustrates the general task flow of DTA. While DTA and DTI share similarities, they differ in aspects such as neural network training focus and task classification - DTA employs regression rather than classification. The evaluation of DTA prediction performance typically utilizes regression metrics such as MSE. Several commonly implemented methods are presented in Fig. 6.

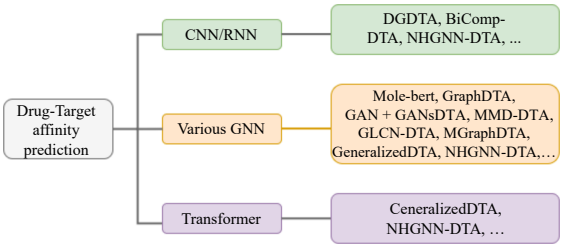

Fig. 6 Common DTA prediction methods.

4.2.1. CNN and RNN-based models for DTA prediction

CNNs and RNNs integrated with fully connected networks demonstrate effectiveness in regression tasks. The drug target affinity prediction network introduced by Zhai et al., combining a dynamic graph attention network and a Bidirectional Long Short-Term Memory (Bi-LSTM) network, exhibited exceptional predictive performance across multiple datasets. The BiComp-DTA network, developed by Kalemati M., Zamani Emani M., and Koohi S., integrates both alignment-free and alignment-based similarity metrics, achieving high effectiveness in modeling complex biological data. Likewise, the attention-based neural network introduced by He et al. adaptively captures salient features from drugs and proteins, demonstrating strong potential for a wide range of drug–target interaction prediction applications [112]. The results are summarized in Table 3.

4.2.2. Various GNNs-based models for DTA prediction

GNNs demonstrate significant effectiveness in processing molecular graph data. The Mole-bert network, introduced by Xia et al., implements masking and variational autoencoder techniques for atom representation pre-training, utilizing DTA prediction tasks as performance metrics, and has demonstrated significant predictive capabilities. The GraphDTA network, developed by Jiang M et al., implements a dual-layer GCN architecture, establishing its reliability across various datasets. The GANsDTA network, created by Zhao et al., implements a semi-supervised GAN approach, demonstrating versatility in handling

Table 3 The results of some representative DTA prediction models.

Model Name	Davis (MSE)	KIBA (MSE)	Model Name	Davis (MSE)	KIBA (MSE)
BiComp-DTA	0.237	0.167	GANsDTA	0.276	0.224
NHGNN-DTA	0.196	0.124	MMD-DTA	0.220	0.134
Mole-bert	0.266	0.157	GLCN-DTA	0.215	0.127
GraphDTA	0.202	0.126	MGraphDTA	0.207	0.128

complex tasks. The MMD-DTA network, developed by Qi et al., utilizes unsupervised learning to identify drug-target binding regions, showing promise in addressing complex biological challenges. The GLCN-DTA network, designed by Qi et al., incorporates graph learning modules into existing architectures, enhancing protein and drug molecular graphs through a soft adjacency matrix. The MGraphDTA network, developed by Yang et al., utilizes ultra-deep densely connected multi-scale GNNs with chemical intuition, achieving significant results across various data environments [113]. The GeneralizedDTA network, featuring a GNN with a dual adaptive mechanism incorporating self-supervision proposed by Lin et al., demonstrated exceptional performance across datasets, particularly in DTA prediction tasks [114]. NHGNN-DTA network similarly employed GNNs [115], showing remarkable adaptability in DTA prediction tasks. Results are presented in Table 3.

4.2.3. Transformer-based models for DTA prediction

The Transformer architecture has been successfully applied to this domain. Both the GeneralizedDTA and NHGNN-DTA networks incorporated the Transformer into their model architectures, achieving significant results. These networks demonstrated robust performance in DTA prediction tasks, particularly in their adaptability across diverse datasets and capability in processing complex data. Results are presented in Table 3.

4.3. DP prediction

DP prediction is a fundamental element of computer-aided drug discovery, aiming to infer the physicochemical and pharmacokinetic characteristics of molecules from intrinsic structural information, such as atomic coordinates and atomic numbers [116]. Key properties include absorption, distribution, metabolism, excretion, and toxicity (ADMET), as well as solubility. Accurate DP prediction provides critical insights into a compound's behavior *in vivo*, thereby informing lead optimization and guiding decision-making in drug development pipelines. Depending on the research objective, studies may focus on predicting individual properties —such as toxicity or solubility —or employ multi-task learning approaches to simultaneously estimate multiple ADMET endpoints. Solubility, which reflects a drug's capacity to dissolve in solvents such as water, is a pivotal factor influencing its pharmacokinetic profile. Early and reliable prediction of these properties facilitates efficient resource allocation by enabling preliminary assessments of drug safety and efficacy.

AI-based prediction tasks typically utilize datasets containing small molecules and their corresponding properties. These molecules undergo encoding through traditional methods or pre-trained models before neural network training. Fully connected networks are frequently employed for binary classification tasks. Fig. 7 illustrates the general prediction workflow. The following section summarizes commonly utilized datasets for Molecular Property Prediction, emphasizing eight toxicity-related datasets from MoleculeNet: Tox21, ToxCast, SIDER, ClinTox, MUV, HIV, BBBP, and BACE [117]. Additional independent datasets focusing on specific properties are included. Table 4 presents detailed dataset information. Recent commonly employed models are discussed and illustrated in Fig. 8.

4.3.1. Various GNNs-based DP prediction

The nodes in GNNs can incorporate various physical and chemical properties of atoms, demonstrating significant effectiveness in molecular property prediction. The Mole-bert network, proposed by Xia et al., integrates a masking strategy with a vari-

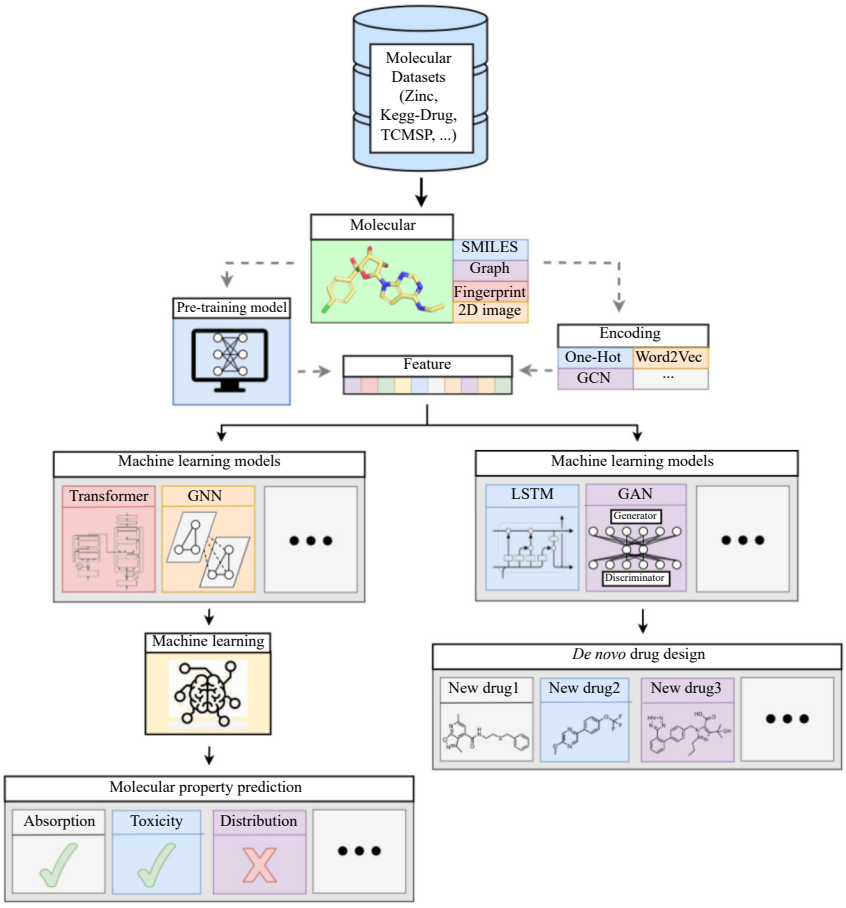

Fig. 7 The common process adopted for drug property prediction and *de novo* drug design.

Table 4 Some datasets used for different drug properties prediction.

Name	Property	Task type	Data size	URL
MUV [117]	Activity	Classification	~ 90 000	https://github.com/deepchem/deepchem/tree/master/datasets
HIV [117]	Activity	Classification	~ 40 000	Same as the one above
BACE [117]	Activity	Classification	~ 1 500	Same as the one above
BBBP [117]	Distribution	Classification	~ 2 000	Same as the one above
Tox21 [117]	Toxicity	Classification	~ 8 000	Same as the one above
ToxCast [117]	Toxicity	Classification	~ 8 000	Same as the one above
SIDER [117]	Toxicity	Classification	~ 1 400	Same as the one above
ClinTox [117]	Toxicity	Classification	~ 1 500	Same as the one above
PAMPA [120]	Absorption	Classification	~ 2 000	https://tdcommons.ai/single_pred_tasks/adme/
HIA [121]	Absorption	Classification	~ 500	https://tdcommons.ai/single_pred_tasks/adme/
Pgp [122]	Absorption	Classification	~ 1 200	https://tdcommons.ai/single_pred_tasks/adme/
Bioavailability [123]	Absorption	Classification	~ 600	https://tdcommons.ai/single_pred_tasks/adme/
Caco-2 [124]	Absorption	Regression	~ 900	https://tdcommons.ai/single_pred_tasks/adme/
Lipophilicity [125]	Absorption	Regression	~ 4 200	https://tdcommons.ai/single_pred_tasks/adme/
Solubility [126]	Absorption	Regression	~ 10 000	https://tdcommons.ai/single_pred_tasks/adme/
Hydration Free Energy [28]	Absorption	Regression	~ 600	https://tdcommons.ai/single_pred_tasks/adme/
PPBR [125]	Distribution	Regression	~ 1 600	https://tdcommons.ai/single_pred_tasks/adme/
VDss [127]	Distribution	Regression	~ 1 100	https://tdcommons.ai/single_pred_tasks/adme/
CYP P450 [128]	Metabolism	Classification	~ 12 000	https://tdcommons.ai/single_pred_tasks/adme/
CYP2C9 [129]	Metabolism	Classification	~ 600	https://tdcommons.ai/single_pred_tasks/adme/
Half Life [130]	Excretion	Regression	~ 600	https://tdcommons.ai/single_pred_tasks/adme/
Clearance [125]	Excretion	Regression	~ 1 000	https://tdcommons.ai/single_pred_tasks/adme/

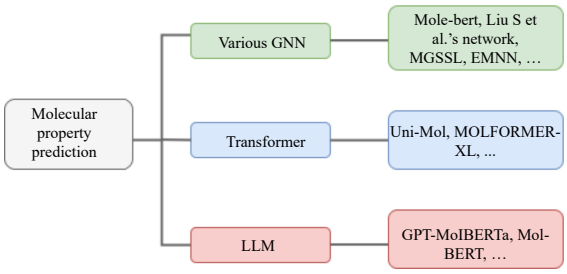

Fig. 8 Common Molecular Property Prediction models.

ational autoencoder for atom-level representation pretraining, delivering superior predictive performance across diverse benchmark datasets. Similarly, the neural network introduced by Liu et al., which utilizes natural language to predict and optimize molecules, exhibited remarkable capabilities, particularly in adapting to diverse datasets. Additionally, the graph self-supervised learning network proposed by Zhang et al., incorporating an innovative base sequence self-supervised framework, demonstrated significant potential in processing complex datasets. Both MGSSL (DFS) and MGSSL (BFS) networks effectively validated their applicability across various datasets [118]. In the message-passing GNNs based on attention and edge memory proposed by Withnall et al., the EMNN network demonstrated robust performance across multiple datasets, particularly excelling in predictive tasks involving the HIV, BBBP, Tox21, and SIDER datasets, illustrating the method's efficiency and adaptability [119].

4.3.2. Transformer-based DP prediction

The attention mechanism in the transformer demonstrates significant effectiveness in molecular property prediction tasks. For instance, in the universal three-dimensional MRL framework proposed by Zhou et al., the deep neural network composed of two identical transformers demonstrated outstanding prediction capabilities, particularly in drug-target prediction tasks across multiple datasets, effectively showcasing the model's adaptability and robustness [98]. In the neural network of the transformer encoder based on the linear attention mechanism and highly distributed training proposed by Ross et al., the MOLFORMER-XL network demonstrated robust predictive capabilities across multiple datasets, particularly excelling in drug-target tasks involving the BBBP, Tox21, ClinTox, HIV, BACE, and SIDER datasets. The approach validated the model's efficiency and adaptability [95].

4.3.3. LLM-based DP prediction

LLMs have gained significant prominence and are being applied across numerous domains, with drug development emerging as a key application area. The GPT-MolBERT network, proposed by Balaji et al., which utilizes GPT for molecular property prediction, demonstrated exceptional performance across multiple datasets, validating its effectiveness in molecular property prediction. Similarly, the Mol-BERT network, proposed by Li et al., which fine-tunes the pre-trained Bert model, exhibited outstanding performance on datasets including BBBP, Tox21, SIDER, and ClinTox, demonstrating the model's potential in diverse

applications.

4.4. *De novo* drug design

De novo, meaning "from the beginning," refers to *de novo* drug design, which encompasses the process of generating novel molecules that meet specific requirements, typically based on biological target information, without a template. This represents a crucial area in contemporary drug development. Traditional methods, such as high-throughput screening, require substantial resources while identifying only a limited number of suitable molecules [131]. The integration of ML techniques has markedly advanced both the efficiency and accuracy of drug design. State-of-the-art generative frameworks, including GANs, VAEs, and diffusion models, can autonomously learn underlying chemical principles while producing structurally novel, synthetically feasible compounds. These models effectively incorporate three-dimensional target structures (such as AlphaFold) [132], pharmacophore models, or gene expression data to achieve conditional molecular generation.

A significant challenge in *de novo* drug design involves the vast number of drug-like molecules in chemical space, complicating extensive exploration. From a data perspective, there exists a critical shortage of high-quality target-ligand interaction data, particularly for historically "undruggable" targets. An additional challenge concerns the synthesizability and physicochemical properties of newly generated molecules. During the generative process, certain synthesized molecular structures exhibit either invalid configurations or present significant synthetic challenges. ML approaches provide transformative advantages for *de novo* drug design, enabling exploration of larger chemical spaces while offering enhanced efficiency and cost-effectiveness compared to traditional methods [133]. The primary distinction between this task and previously mentioned tasks lies in neural network selection. It requires neural networks capable of generation tasks, such as LSTM, GANs, Transformer, etc. The general task flow of *de novo* drug design is shown in Fig. 6. The following section presents several commonly used methods as illustrated in Fig. 9.

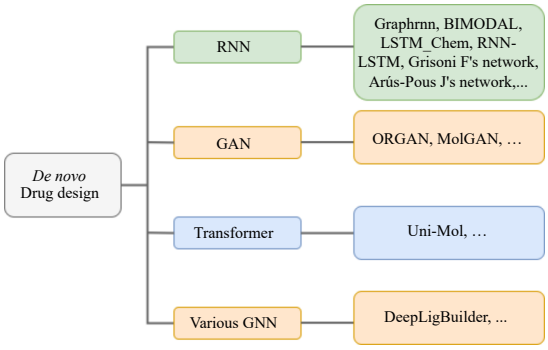

Fig. 9 Common *De novo* Drug design models.

4.4.1. RNN-based generation models

RNNs demonstrate effectiveness in sequence data generation, making them applicable to *de novo* drug design. You et al. developed a graph generation model utilizing RNNs for node and edge generation [134]. Grisoni et al. introduced a SMILES generation network incorporating bidirectional LSTM with three bidirectional strategies [135]. Zhang et al. developed a generation network integrating LSTM and dense full CNNs [54]. Zhang et al. subsequently created LSTM_Pep, a model capable of generating *de novo* peptides [136]. Through fine-tuning, this model generates *de novo* peptides with specific therapeutic properties. Gupta et al. implemented a generation network utilizing LSTM for recursive drug generation. Grisoni et al. developed an LSTM pre-trained and fine-tuned using LXRα agonists [137]. Arús-Pous et al. introduced an RNN based on data augmentation through acyclic bond splitting [138]. RNNs offer significant advantages for *de novo* drug design, particularly in processing sequential molecular representations like SMILES, while requiring modest computational resources. However, these networks exhibit limitations in modeling long-range molecular interactions, often producing invalid structures for complex molecules. Additionally, RNNs lack the capacity to incorporate essential 2D/3D spatial information necessary for accurate stereochemical representation.

4.4.2. GAN-based generation models

GANs generate desired molecules through encoder-decoder adversarial processes, demonstrating effectiveness in *de novo* drug design tasks. Guimaraes et al. developed a generation network integrating reinforcement learning and GANs to optimize sequence distribution [139]. De Cao N, Kipf T introduced a GAN molecular generation network incorporating GCNs [140]. Wang et al. [141] developed an integrated molecular generation approach combining GANs with LSTM architectures. GANs excel in generating chemically valid and structurally novel molecules, typically producing greater diversity than RNN-based methods. Their capabilities extend to direct molecular graph or 3D structure generation, surpassing sequence-based models. However, GANs require substantial high-quality training data for discriminator development, with limited datasets susceptible to overfitting or invalid molecular generation. Furthermore, generated structures may violate chemical valence rules, necessitating post-processing or constrained loss functions. These challenges contribute to increased training complexity compared to conventional approaches.

4.4.3. LLM-based generation models

Recent advancements in natural language processing demonstrate how utilizing unlabeled background information enhances performance in general language tasks, highlighting current model capabilities. Similarly, integrating comprehensive chemical knowledge into models potentially advances drug discovery significantly. Transformers function as generative models, achieving notable results in *de novo* drug design tasks. Zhou et al. developed a deep neural network comprising two identical Transformers within a universal three-dimensional MRL framework. The self-attention mechanism effectively addresses long-range dependencies in molecular sequences, overcoming RNN limitations in modeling extended structural patterns. Furthermore, Transformer's parallel computation capabilities enhance generation efficiency. Transformer, an encoder-decoder-based model, enables direct *de novo* drug generation and molecular feature extraction from large datasets. The decoder-based GPT model finds widespread application in *de novo* drug design, exemplified by cMolGPT [142], MolGPT [143], and SMILES GPT [144]. While LLMs advance small MG, optimal performance requires extensive molecular dataset pre-training to avoid overfitting. Additionally, LLM-based generation models typically face challenges in incorporating 3D structural information, being primarily sequence-based or 2D graph-based models.

4.4.4. Various GNN-based generation models

GNNs, when combined with complementary machine learning algorithms, have shown considerable effectiveness in molecular generation tasks. Li et al. introduced DeepLigBuilder, which integrates a MPNN-based ligand generation model with Monte Carlo tree search [145]. By directly processing graph-structured molecular data, GNNs inherently reduce the information loss often associated with linearized representations such as SMILES. Their message-passing mechanisms enable accurate modeling of atomic interactions, capturing both local chemical environments and the global molecular topology. The incorporation of chemical con-

straints during the generation process further minimizes the likelihood of producing invalid molecular structures. However, most GNN-based generative frameworks adopt a sequential atom–bond addition strategy, which is generally slower than parallelized generation approaches employed by models such as Transformers or GANs. Furthermore, these methods typically require extensive high-quality molecular graph datasets to achieve optimal generalization and performance.

5. Conclusions

Despite the promising potential of integrating molecular representations with deep learning in drug development, several critical challenges and future research directions require attention to advance the field.

First, existing molecular representation methods-such as SMILES strings, molecular graphs, and 3D grids —encounter a fundamental trade-off between information retention and computational efficiency. For instance, RNNs are commonly employed for processing sequential information. While SMILES provides a compact and sequence-based representation, it lacks essential spatial and stereochemical information. CNNs and GNNs demonstrate particular effectiveness in processing structured information. 3D grids deliver comprehensive structural details but incur high computational costs, limiting their practicality for large-scale applications. To address this limitation, future research should prioritize developing hybrid representations that maintain molecular fidelity while remaining computationally feasible. Additionally, developing higher-quality pre-trained molecular models for feature extraction could substantially enhance efficiency without compromising accuracy.

Second, the limited availability of high-quality, well-annotated molecular data continues to present a significant challenge. Experimental noise, dataset biases, and the substantial cost of obtaining reliable biological annotations restrict model generalization. To address these challenges, techniques such as few-shot learning, self-supervised pre-training, and data augmentation could enhance data efficiency. Moreover, training large-scale foundation models on diverse, extensive datasets may improve generalization across various drug discovery tasks, representing a crucial direction for future research.

Third, the opaque decision-making processes of deep neural networks present substantial interpretability challenges in drug discovery, where elucidating the basis of model outputs is vital for clinical integration. The adoption of explainable AI methodologies, such as attention mechanisms, saliency maps, and graph-based interpretability frameworks, is essential for enhancing transparency, rigorously validating predictions, and building confidence among researchers, clinicians, and regulatory bodies.

Finally, while current models demonstrate excellence in single-task settings, they frequently encounter difficulties in generalizing to novel targets, unseen diseases, or multi-objective optimization scenarios. This limitation emphasizes the necessity for universal molecular pretrained models-akin to LLMs in NLP-that can transfer knowledge across diverse tasks. While molecular LLMs have significantly advanced drug discovery, important limitations persist. Current biological and chemical LLMs struggle to explicitly incorporate 3D structural information, which is often more critical than sequence data alone for accurately characterizing molecular functions and properties. Structural data — typically represented as atomic coordinates—does not integrate seamlessly into conventional language modeling architectures. Recent research has explored encoding 3D structures as learnable "structural tokens" alongside sequence tokens; however, these approaches remain in the early stages of development. Furthermore, multi-task learning frameworks that jointly optimize related molecular objectives could improve both robustness and translational applicability. Future deep learning approaches for molecular tasks should adopt multimodal representations that integrate structural, physicochemical, and biological information to maximize data completeness and relevance. Leveraging high-performance pretrained large models may further enhance generalization, while incorporating diverse biological indicators into optimization objectives can strengthen predictive and generative performance. Finally, ensuring sufficient interpretability is essential for enabling human understanding and fostering trust in model-driven decisions.

Funding

This work is supported by the Shenzhen Key Laboratory of Intelligent Bioinformatics (No. ZDSYS20220422103800001), the Shenzhen Science and Technology Program (No. JCYJ20230807140709020), National Natural Science Foundation of China (Nos. 62402489, U22A2041, and 62373172), and the China Postdoctoral Science Foundation (No. 2023M743688), Guangdong Basic and Applied Basic Research Foundation (Nos. 2024A1515011960 and 2023A1515110570).

Declaration of competing interest

These authors have no conflict of interest to declare.

References

1 Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3(8):711-716.https://doi.org/10.1038/nrd1470.
2 Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. 2010;9(3):203-214. https://doi.org/10.1038/nrd3078.
3 Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov*. 2004;3(5):417-429. https://doi.org/10.1038/nrd1382.
4 Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. 2019;20(2):273-286. https://doi.org/10.1093/biostatistics/kxy072.
5 Omejc M. Drug development: the journey of a medicine from lab to shelf. *J Dev Drugs*. 2020;9(1):e115. https://doi.org/10.35248/2329-6631.20.9.e155.
6 Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res*. 2024;52(D1):D1265-D1275. https://doi.org/10.1093/nar/gkad976.
7 Kim S, Chen J, Cheng T, et al. PubChem 2023 update. *Nucleic Acids Res*. 2023;51(D1):D1373-D1380. https://doi.org/10.1093/nar/gkac956.
8 Hearst MA, Dumais ST, Osuna E, et al. Support vector machines. *IEEE Intell Syst*. 1998;13(4):18-28. https://doi.org/10.1109/5254.708428.
9 Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. https://doi.org/10.1023/A:1010933404324.
10 Albawi S, Mohammed TA, AlZawi S, et al. Understanding of a convolutional neural network. *ICET'17*. 2017;1-6.https://doi.org/10.1109/ICEngTechnol.2017.8308186.
11 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-80. https://doi.org/10.1162/neco.1997.9.8.1735.
12 Devlin J, Chang MW, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2018;04805. https://doi.org/10.18653/v1/N19-1423.
13 Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach*. 2020;30:681-694. https://doi.org/10.1007/s11023-020-09548-1.
14 Mouchlis VD, Melagraki G, Zacharia LC, et al. Computer-aided drug design of β-secretase, γ-secretase and anti-tau inhibitors for the discovery of novel Alzheimer's therapeutics. *Int J Mol Sci*. 2020;21(3):703. https://doi.org/10.3390/ijms21030703.
15 Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? *J Chem Inf Model*. 2012;52(6):1413-1437. https://doi.org/10.1021/ci200409x.
16 Favre HA, Powell WH. International union of pure and applied chemistry. Nomenclature of organic chemistry: IUPAC recommendations and preferred names 2013. London, LD: LRSC Publishing; 2013.
17 Weininger D. SMILES, a chemical language and information system.1. Introduction to methodology and encoding rules. *Chem Inf Comput Sci*. 1988;28(1):31-36. https://doi.org/10.1021/ci00057a005.
18 Kearnes S, McCloskey K, Berndl M, et al. Molecular graph convolutions: moving beyond fingerprints. *J Comput-Aided Mol Des*. 2016;30:595-608. https://doi.org/10.1007/s10822-016-9938-8.
19 Lawlor B. The chemical structure association trust: advancing scientific discovery for fifty years. *Chem Int*. 2016;38(2):12-15. https://doi.org/10.1515/ci-2016-0206.

20 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630(8016):493-500. https://doi.org/10.1038/s41586-024-07487-w.
21 He X, Li J, Shen S, et al. AlphaFold3 versus experimental structures: assessment of the accuracy in ligand-bound G protein-coupled receptors. *Acta Pharmacol Sin*. 2025;46(4):1111-1122. https://doi.org/10.1038/s41401-024-01429-y.
22 Tang X, Dai H, Knight E, et al. A survey of generative AI for *de novo* drug design: new frontiers in molecule and protein generation. *Briefings Bioinf*. 2024;25(4):bbae338. https://doi.org/10.1093/bib/bbae338.
23 Suruliandi A, Idhaya T, Raja S. Drug target interaction prediction using machine learning techniques–a review. *Int J Interact Multimedia Artif Intell*. 2024;8(6):86-100. https://doi.org/10.9781/ijimai.2022.11.002.
24 Zeng X, Li S, Lv S, et al. A comprehensive review of the recent advances on predicting drug-target affinity based on deep learning. *Front Pharmacol*. 2024;15:1375522. https://doi.org/10.3389/fphar.2024.1375522.
25 Chen M, Jiang Y, Lei X, et al. Drug-target interactions prediction based on signed heterogeneous graph neural networks. *Chin J Electron*. 2024;33(1):231-244. https://doi.org/10.23919/cje.2022.00.384.
26 Singh S, Kaur N, Gehlot A. Application of artificial intelligence in drug design: a review. *Comput Biol Med*. 2024;179:108810. https://doi.org/10.1016/j.compbiomed.2024.108810.
27 Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. *Nat Biotechnol*. 2020;38(2):143-145. https://doi.org/10.1038/s41587-020-0418-2.
28 Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J Comput-Aided Mol Des*. 2014;28:711-720. https://doi.org/10.1007/s10822-014-9747-x.
29 Simian C, Binxin D, Zhang D, et al. Advances in intelligent mass spectrometry data processing technology for *in vivo* analysis of natural medicines. *Chin J Nat Med*. 2024;22(10):900-913. https://doi.org/10.1016/S1875-5364(24)60687-4.
30 Li Shao, Xiao Wei. General expert consensus on the application of network pharmacology in the research and development of new traditional Chinese medicine drugs. *Chin J Nat Med*. 2025;23(2):129-142. http://10.1016/S1875-5364(25)60802-8.
31 Cannon M, Stevenson J, Stahl K, et al. DGIdb 5. 0: rebuilding the drug–gene interaction database for precision medicine and drug discovery platforms. *Nucleic Acids Res*. 2024;52(D1):D1227-D1235. https://doi.org/10.1093/nar/gkad1040.
32 Harding SD, Armstrong JF, Faccenda E, et al. The IUPHAR/BPS guide to Pharmacology in 2024. *Nucleic Acids Res*. 2024;52(D1):D1438-D1449. https://doi.org/10.1093/nar/gkad944.
33 Avram S, Wilson TB, Curpan R, et al. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Res*. 2023;51(D1):D1276-D1287. https://doi.org/10.1093/nar/gkac1085.
34 Zhou Y, Zhang Y, Zhao D, et al. TTD: therapeutic target database describing target druggability information. *Nucleic Acids Res*. 2024;52(D1):D1465-D77. https://doi.org/10.1093/nar/gkad751.
35 Tingle BI, Tang KG, Castanon M, et al. ZINC-22— A free multi-billion-scale database of tangible compounds for ligand discovery. *J Chem Inf Model*. 2023;63(4):1166-76. https://doi.org/10.1021/acs.jcim.2c01253.
36 Gallo K, Goede A, Eckert A, et al. PROMISCUOUS 2.0: a resource for drug-repositioning . *Nucleic Acids Res*. 2021;49(D1):D1373-D1380. https://doi.org/10.1093/nar/gkaa1061.
37 Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;45(D1):D945-D954. https://doi.org/10.1093/nar/gkw1074.
38 Siramshetty VB, Eckert OA, Gohlke BO, et al. SuperDRUG2: a one stop resource for approved/marketed drugs. *Nucleic Acids Res*. 2018;46(D1):D1137-D1143. https://doi.org/10.1093/nar/gkx1088.
39 Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44(D1):D1075-D1079. https://doi.org/10.1093/nar/gkv1075.
40 Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1):D353-D361. https://doi.org/10.1093/nar/gkw1092.
41 Szklarczyk D, Santos A, Von Mering C, et al. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res*. 2016;44(D1):D380-D384. https://doi.org/10.1093/nar/gkv1277.
42 Gilson MK, Liu T, Baitaluk M, et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. 2016;44(D1):D1045-D1053. https://doi.org/10.1093/nar/gkv1072.
43 Ru J, Li P, Wang J, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J Cheminf*. 2014;6:1-6. https://doi.org/10.1186/1758-2946-6-13.
44 Kumar R, Chaudhary K, Gupta S, et al. CancerDR: cancer drug resistance database. *Sci Rep*. 2013;3(1):1445. https://doi.org/10.1038/srep01445.
45 Hecker N, Ahmed J, Von Eichborn J, et al. SuperTarget goes quantitative: update on drug–target interactions. *Nucleic Acids Res*. 2012;40(D1):D1113-D1117. https://doi.org/10.1093/nar/gkr912.
46 Xu Y, Lin K, Wang S, et al. Deep learning for molecular generation. *Future Med Chem*. 2019;11(6):567-597. https://doi.org/10.4155/fmc-2018-0358.
47 Carpenter KA, Cohen DS, Jarrell JT, et al. Deep learning and virtual drug screening. *Future Med Chem*. 2018;10(21):2557-2567. https://doi.org/10.4155/fmc-2018-0314.
48 O'Boyle NM. Towards a universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *J Cheminf*. 2012;4:1-14. https://doi.org/10.1186/1758-2946-4-22.
49 Homer RW, Swanson J, Jilek RJ, et al. SYBYL line notation (SLN): a single notation to represent chemical structures, queries, reactions, and virtual libraries. *J Chem Inf Model*. 2008;48(12):2294-307. https://doi.org/10.1021/ci7004687.
50 Atz K, Cotos L, Isert C, et al. Prospective *de novo* drug design with deep interactome learning. *Nat Commun*. 2024;15(1):3408. https://doi.org/10.1038/s41467-024-47613-w.
51 Yang J, Cai Y, Zhao K, et al. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov Today*. 2022;27(11):103356. https://doi.org/10.1016/j.drudis.2022.103356.
52 Rarey M, Dixon JS. Feature trees: a new molecular similarity measure based on tree matching. *J Comput-Aided Mol Des*. 1998;12:471-490. https://doi.org/10.1023/a:1008068904628.
53 Rong X. word2vec parameter learning explained. *arXiv*. 2014;14112738. https://doi.org/10.48550/arXiv.1411.2738
54 Zhang H, Saravanan KM, Yang Y, et al. Generating and screening *de novo* compounds against given targets using ultrafast deep learning models as core components. *Briefings Bioinf*. 2022;23(4):bbac226. https://doi.org/10.1093/bib/bbac226.
55 Gupta A, Müller AT, Huisman BJ, et al. Generative recurrent networks for *de novo* drug design. *Mol Inf*. 2018;37(1-2):1700111. https://doi.org/10.1002/minf.201700111.
56 Wu C, Zhang X, Yang Z, et al. Learning to SMILES: ban-based strategies to improve latent representation learning from molecules. *Briefings Bioinf*. 2021;22(6):bbab327. https://doi.org/10.1093/bib/bbab327.
57 Wang Y, You Z, Yang S, et al. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Med Inf Decis Making*. 2020;20:1-9. https://doi.org/10.1186/s12911-020-1052-0.
58 Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. *EMNLP'2020*. 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6.
59 Huang K, Xiao C, Glass LM, et al. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*. 2021;37(6):830-836. https://doi.org/10.1093/bioinformatics/btaa880.
60 Balaji S, Magar R, Jadhav Y, et al. Gpt-molberta: Gpt molecular features language model for molecular property prediction. *arXiv*. 2023;231003030. https://doi.org/10.48550/arXiv.2310.03030.
61 Kalemati M, Zamani Emani M, Koohi S. BiComp-DTA: drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach. *PLoS Comput Biol*. 2023;19(3):e1011036. https://doi.org/10.1371/journal.pcbi.1011036.
62 Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview. *IEEE Signal Process Mag*. 2018;35(1):53-65. https://doi.org/10.1109/MSP.2017.2765202.
63 Zhao L, Wang J, Pang L, et al. GANsDTA: predicting drug-target binding affinity using GANs. *Front Genet*. 2020;10:1243. https://doi.org/10.3389/fgene.2019.01243.
64 Chen Z, You Z, Guo Z, et al. Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model. *Front Bioeng Biotechnol*. 2020;8:338. https://doi.org/10.3389/fbioe.2020.00338.
65 Scarselli F, Gori M, Tsoi AC, et al. The graph neural network model. *IEEE Trans Neural Networks*. 2008;20(1):61-80. https://doi.org/10.1109/TNN.2008.2005605.
66 Jiang M, Li Z, Zhang S, et al. Drug–target affinity prediction using graph neural network and contact maps. *RSC Adv*. 2020;10(35):20701-20712. https://doi.org/10.1039/D0RA02297G.
67 Zhai H, Hou H, Luo J, et al. DGDTA: dynamic graph attention network for predicting drug–target binding affinity. *BMC Bioinf*. 2023;24(1):367. https://doi.org/10.1186/s12859-023-05497-5.
68 Nguyen T, Le H, Quinn TP, et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*. 2021;37(8):1140-1147. https://doi.org/10.1093/bioinformatics/btaa921.
69 Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. 6th International Conference on Learning Representations Iclr 2018 Conference Track Proceedings. 2017;1050(20):48550.https://doi.org/10.17863/CAM.48429.
70 Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks? *arXiv*. 2018;1810.00826.https://doi.org/10.48550/arXiv.1810.00826
71 Zhang S, Tong H, Xu J, et al. Graph convolutional networks: a comprehensive review. *Comput Soc Netw*. 2019;6(1):1-23. https://doi.org/10.1186/s40649-019-0069-y.
72 Qi H, Yu T, Yu W, et al. Drug–target affinity prediction with extended graph learning-convolutional networks. *BMC Bioinf*. 2024;25(1):75. https://doi.org/10.1186/s12859-024-05698-6.
73 Bai P, Miljković F, John B, et al. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat Mach Intell*. 2023;5(2):126-136. https://doi.org/10.1038/s42256-022-00605-1.
74 Fang X, Liu L, Lei J, et al. Geometry-enhanced molecular representation learning for property prediction. *Nat Mach Intell*. 2022;4(2):127-134. https://doi.org/10.1038/s42256-021-00438-4.
75 Xia J, Zhao C, Hu B, et al. Mole-bert: rethinking pre-training graph neural networks for molecules. *ChemRxiv*. 2023. https://doi.org/10.26434/chemrxiv-2023-dngg4.
76 Rifaioglu AS, Nalbat E, Atalay V, et al. DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem Sci*. 2020;11(9):2531-2557. https://doi.org/10.1039/C9SC03414E.
77 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(11):1004-1016. https://doi.org/10.1038/s42256-022-00557-6.
78 Casey AD, Son SF, Bilionis I, et al. Prediction of energetic material properties from electronic structure using 3D convolutional neural

networks. *J Chem Inf Model*. 2020;60(10):4457-4573. https://doi.org/10.1021/acs.jcim.0c00259.
79 Kuzminykh D, Polykovskiy D, Kadurin A, et al. 3D molecular representations based on the wave transform for convolutional neural networks. *Mol Pharmaceutics*. 2018;15(10):4378-4385. https://doi.org/10.1021/acs.molpharmaceut.7b01134.
80 Wang Z, Liu M, Luo Y, et al. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. *Bioinformatics*. 2022;38(9):2579-2586. https://doi.org/10.1093/bioinformatics/btac112.
81 Clevert DA, Le T, Winter R, et al. Img2Mol–accurate SMILES recognition from molecular graphical depictions. *Chem Sci*. 2021;12(42):14174-14181. https://doi.org/10.1039/D1SC01839F.
82 Li J, Jiang X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. *Wirel Commun Mob Comput*. 2021;2021(1): 7181815. https://doi.org/10.1155/2021/7181815.
83 Stärk H, Beaini D, Corso G, et al. 3d infomax improves gnns for molecular property prediction. *arXiv*. 2022;2110.04126.https://doi.org/10.48550/arXiv.2110.04126.
84 Xia X, Zhu C, Zhong F, et al. MDTips: a multimodal-data-based drug–target interaction prediction system fusing knowledge, gene expression profile, and structural data. *Bioinformatics*. 2023;39(7):btad411. https://doi.org/10.1093/bioinformatics/btad411.
85 Liu S, Nie W, Wang C, et al. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nat Mach Intell*. 2023;5(12):1447-1457. https://doi.org/10.1038/s42256-023-00759-6.
86 Edwards C, Zhai C, Ji H, et al. Text2mol: cross-modal molecule retrieval with natural language queries. EMNLP'2021. 2021.https://doi.org/10.18653/v1/2021.emnlp-main.47.
87 Zhang Q, Wei Y, Liao B, et al. MMD-DTA: a multi-modal deep learning framework for drug-target binding affinity and binding region prediction. *IEEE ACM T COMPUT BI*. 2024;21(6):2200-2211. https://doi.org/10.1109/TCBB.2024.3451985.
88 Lu X, Xie L, Xu L, et al. Multimodal fused deep learning for drug property prediction: integrating chemical language and molecular graph. *Comput Struct Biotechnol J*. 2024;23:1666-1679. https://doi.org/10.1016/j.csbj.2024.04.030.
89 Xie L, Xu L, Kong R, et al. Improvement of prediction performance with conjoint molecular fingerprint in deep learning. *Front Pharmacol*. 2020;11: 606668. https://doi.org/10.3389/fphar.2020.606668.
90 Meng Z, Chen C, Zhang X, et al. Exploring fragment adding strategies to enhance molecule pretraining in AI-driven drug discovery. *Big Data Min Anal*. 2024;7(3):565-576. https://doi.org/10.26599/BDMA.2024.9020003.
91 Kim D, Lee W, Hwang SJ. Mol-LLaMA: towards general understanding of molecules in large molecular language model. *arXiv*. 2025; 250213449. https://doi.org/10.48550/arXiv.2502.13449.
92 Lee C, Song Y, Jeong Y, et al. Mol-LLM: generalist molecular LLM with improved graph utilization. *arXiv*. 2025; 250202810.https://doi.org/10.48550/arXiv.2502.02810.
93 Ahmad W, Simon E, Chithrananda S, et al. Chemberta-2: towards chemical foundation models. *arXiv*. 2022; 220901712.https://doi.org/10.48550/arXiv.2209.01712.
94 AbdelAty H, Gould IR. Large-scale distributed training of transformers for chemical fingerprinting. *J Chem Inf Model*. 2022;62(20):4852-4862. https://doi.org/10.1021/acs.jcim.2c00715.
95 Ross J, Belgodere B, Chenthamarakshan V, et al. Large-scale chemical language representations capture molecular structure and properties. *Nat Mach Intell*. 2022;4(12):1256-1264. https://doi.org/10.1038/s42256-022-00580-7.
96 Irwin R, Dimitriadis S, He J, et al. Chemformer: a pre-trained transformer for computational chemistry. *Mach Learn: Sci Technol*. 2022;3(1):015022. https://doi.org/10.1088/2632-2153/ac3ffb.
97 Xue D, Zhang H, Xiao D, et al. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv*. 2020; 12. 23. 424259.https://doi.org/10.1016/j.scib.2022.01.029.
98 Zhou G, Gao Z, Ding Q, et al. Uni-mol: a universal 3d molecular representation learning framework. *ChemRxiv*. 2023.https://doi.org/10.26434/chemrxiv-2022-jjm0j.
99 Toniato A, Vaucher AC, Schwaller P, et al. Enhancing diversity in language based models for single-step retrosynthesis. *Digit Discov*. 2023;2(2):489-501. https://doi.org/10.1039/d2dd00110a.
100 Yüksel A, Ulusoy E, Ünlü A, et al. SELFormer: molecular representation learning *via* SELFIES language models. *Mach Learn: Sci Technol*. 2023;4(2): 025035. https://doi.org/10.1088/2632-2153/acdb30.
101 Cho KH, No KT. iupacGPT: IUPAC-based large-scale molecular pre-trained model for property prediction and molecule generation. *Mol Divers*. 2025; 2025: 1-9.https://doi.org/10.1007/s11030-025-11280-w.
102 Liu Y, Ding S, Zhou S, et al. Moleculargpt: open large language model (llm) for few-shot molecular property prediction. *arXiv*. 2024; 240612950.https://doi.org/10.48550/arXiv.2406.12950.
103 Bagherian M, Sabeti E, Wang K, et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings Bioinf*. 2021;22(1):247-269. https://doi.org/10.1093/bib/bbz157.
104 An Q, Yu L. A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Briefings Bioinf*. 2021;22(6): bbab275. https://doi.org/10.1093/bib/bbab275.
105 Yu L, Qiu W, Lin W, et al. HGDTI: predicting drug–target interaction by using information aggregation based on heterogeneous graph neural network. *BMC Bioinf*. 2022;23(1):126. https://doi.org/10.1186/s12859-022-04655-5.
106 Li Y, Qiao G, Wang K, et al. Drug–target interaction predication *via* multi-channel graph neural networks. *Briefings Bioinf*. 2022;23(1):bbab346. https://doi.org/10.1093/bib/bbab346.
107 Cheng Z, Yan C, Wu F, et al. Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE ACM T COMPUT BI*. 2021;19(4):2208-18. https://doi.org/10.1109/TCBB.2021.3077905.
108 Zhang Y, Hu Y, Han N, et al. A survey of drug-target interaction and affinity prediction methods *via* graph neural networks. *Comput Biol Med*. 2023;163: 107136. https://doi.org/10.1016/j.compbiomed.2023.107136.
109 Lu Z, Song G, Zhu H, et al. DTIAM: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nat Commun*. 2025;16(1):2548. https://doi.org/10.1038/s41467-025-57828-0.
110 Wang N, Li P, Hu X, et al. Herb target prediction based on representation learning of symptom related heterogeneous network. *Comput Struct Biotechnol J*. 2019;17:282-290. https://doi.org/10.1016/j.csbj.2019.02.002.
111 Duan P, Yang K, Su X, et al. HTINet2: herb–target prediction *via* knowledge graph embedding and residual-like graph neural network. *Briefings Bioinf*. 2024;25(5):bbae414. https://doi.org/10.1093/bib/bbae414.
112 He H, Chen G, Chen CYC. NHGNN-DTA: a node-adaptive hybrid graph neural network for interpretable drug–target binding affinity prediction. *Bioinformatics*. 2023;39(6):btad355. https://doi.org/10.1093/bioinformatics/btad355.
113 Yang Z, Zhong W, Zhao L, et al. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chem Sci*. 2022;13(3):816-33. https://doi.org/10.1039/D1SC05180F.
114 Lin S, Shi C, Chen J. GeneralizedDTA: combining pre-training and multi-task learning to predict drug-target binding affinity for unknown drug discovery. *BMC Bioinf*. 2022;23(1):367. https://doi.org/10.1186/s12859-022-04905-6.
115 Liu T, Yang X, Zhou H, et al. A survey of collaborative filtering recommender algorithms based on graph neural networks. *J Integr Technol*. 2024;13(4):1-15. https://doi.org/10.12146/j.issn.2095-3135.20230731001.
116 Kuang T, Liu P, Ren Z. Impact of domain knowledge and multi-modality on intelligent molecular property prediction: a systematic survey. *Big Data Min Anal*. 2024;7(3):858-88. https://doi.org/10.26599/bdma.2024.9020028.
117 Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 2018;9(2):513-30. https://doi.org/10.1039/C7SC02664A.
118 Zhang Z, Liu Q, Wang H, et al. Motif-based graph self-supervised learning for molecular property prediction. *Adv Neural Inf Process Syst*. 2021;34: 15870-15882. https://doi.org/10.48550/arXiv.2110.00987.
119 Withnall M, Lindelöf E, Engkvist O, et al. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *J Cheminf*. 2020;12(1):1. https://doi.org/10.1186/s13321-019-0407-y.
120 Siramshetty V, Williams J, Nguyễn DT, et al. Validating ADME QSAR models using marketed drugs. *Slas Discov*. 2021;26(10):1326-1336. https://doi.org/10.1177/24725552211017520.
121 Hou T, Wang J, Zhang W, et al. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J Chem Inf Model* . 2007;47(1):208-218. https://doi.org/10.1021/ci600343x.
122 Broccatelli F, Carosati E, Neri A, et al. A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J Med Chem*. 2011;54(6):1740-1751. https://doi.org/10.1021/jm101421d.
123 Ma C, Yang S, Zhang H, et al. Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method. *J Pharm Biomed Anal*. 2008;47(4-5):677-682. https://doi.org/10.1016/j.jpba.2008.03.023.
124 Wang N, Dong J, Deng Y, et al. ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting. *J Chem Inf Model*. 2016;56(4):763-773. https://doi.org/10.1021/acs.jcim.5b00642.
125 Wenlock M, Tomkinson N. Experimental *in vitro* DMPK and physicochemical data on a set of publicly disclosed compound. ChEMBL. 2015. https://doi org/106019/CHEMBL3301361.
126 Sorkun MC, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data*. 2019;6(1):143. https://doi.org/10.1038/s41597-019-0151-1.
127 Lombardo F, Jing Y. *In silico* prediction of volume of distribution in humans. Extensive data set and the exploration of linear and nonlinear methods coupled with molecular interaction fields descriptors. *J Chem Inf Model*. 2016;56(10):2042-2052. https://doi.org/10.1021/acs.jcim.6b00044.
128 Veith H, Southall N, Huang R, et al. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat Biotechnol*. 2009;27(11):1050-1055. https://doi.org/10.1038/nbt.1581.
129 CarbonMangels M, Hutter MC. Selecting relevant descriptors for classification by bayesian estimates: a comparison with decision trees and support vector machines approaches for disparate data sets. *Mol Inform*. 2011;30(10):885-895. https://doi.org/10.1002/minf.201100069.
130 Obach RS, Lombardo F, Waters NJ. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metab Dispos*. 2008;36(7):1385-1405. https://doi.org/10.1124/dmd.118.082966.
131 Xu Z, Lei X, Ma M, et al. Molecular generation and optimization of molecular properties using a transformer model. *Big Data Min Anal*. 2023;7(1):142-155. https://doi.org/10.26599/BDMA.2023.9020009.
132 Xia C, Tang Q. Uncovering the statistical trends of protein evolution with AlphaFold database. *J Integr Technol*. 2023;13(2):74-88. https://doi.org/10.12146/j.issn.2095-3135.20230912001.
133 Mouchlis VD, Afantitis A, Serra A, et al. Advances in *de novo* drug design: from conventional to machine learning methods. *Int J Mol Sci*. 2021;22(4): 1676. https://doi.org/10.3390/ijms22041676.
134 You J, Ying R, Ren X, et al. Graphrnn: generating realistic graphs with deep auto-regressive models. *ICML*. 2018. https://arxiv.org/abs/1802.08773.
135 Grisoni F, Moret M, Lingwood R, et al. Bidirectional molecule generation

with recurrent neural networks. *J Chem Inf Model*. 2020;60(3):1175-1183. https://doi.org/10.1021/acs.jcim.9b00943.
136 Zhang H, Saravanan KM, Wei Y, et al. Deep learning-based bioactive therapeutic peptide generation and screening. *J Chem Inf Model*. 2023;63(3): 835-845. https://doi.org/10.1021/acs.jcim.2c01485.
137 Grisoni F, Huisman BJ, Button AL, et al. Combining generative artificial intelligence and on-chip synthesis for *de novo* drug design. *Sci Adv*. 2021;7 (24):eabg3338. https://doi.org/10.1126/sciadv.abg3338.
138 ArúsPous J, Patronov A, Bjerrum EJ, et al. SMILES-based deep generative scaffold decorator for *de-novo* drug design. *J Cheminf*. 2020;12:1-18. https://doi.org/10.1186/s13321-020-00441-8.
139 Guimaraes GL, SanchezLengeling B, Outeiral C, et al. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv*. 2017; 170510843. https://doi.org/10.48550/arXiv.1705.10843.
140 De Cao N, Kipf T. MolGAN: an implicit generative model for small molecular graphs. *arXiv*. 2018; 180511973. https://doi.org/10.48550/arXiv.1805.11973.
141 Wang F, Feng X, Guo X, et al. Improving *de novo* molecule generation by embedding LSTM and attention mechanism in CycleGAN. *Front Genet*. 2021;12:709500. https://doi.org/10.3389/fgene.2021.709500.
142 Wang Y, Zhao H, Sciabola S, et al. cMolGPT: a conditional generative pre-trained transformer for target-specific *de novo* molecular generation. *Molecules*. 2023;28(11):4430. https://doi.org/10.3390/molecules28114430.
143 Bagal V, Aggarwal R, Vinod P, et al. MolGPT: molecular generation using a transformer-decoder model. *J Chem Inf Model*. 2021;62(9):2064-2076. https://doi.org/10.1021/acs.jcim.1c00600.
144 Adilov S. Generative pre-training from molecules. *ChemRxiv*. 2021. https://doi.org/10.26434/chemrxiv-2021-5fwjd.
145 Li Y, Pei J, Lai L. Structure-based *de novo* drug design using 3D deep generative models. *Chem Sci*. 2021;12(41):13664-13675. https://doi.org/10.1039/d1sc04444c.