



REVIEW

Open Access



Bridging chemical space and biological efficacy: advances and challenges in applying generative models in structural modification of natural products

Chuan-Su Liu^{1,2†}, Bing-Chao Yan^{2†}, Han-Dong Sun², Jin-Cai Lu^{1*} and Pema-Tenzin Puno^{2*} 

Abstract

Natural products (NPs) are invaluable resources for drug discovery, characterized by their intricate scaffolds and diverse bioactivities. AI drug discovery & design (AIDD) has emerged as a transformative approach for the rational structural modification of NPs. This review examines a variety of molecular generation models since 2020, focusing on their potential applications in two primary scenarios of NPs structure modification: modifications when the target is identified and when it remains unidentified. Most of the molecular generative models discussed herein are open-source, and their applicability across different domains and technical feasibility have been evaluated. This evaluation was accomplished by integrating a limited number of research cases and successful practices observed in the molecular optimization of synthetic compounds. Furthermore, the challenges and prospects of employing molecular generation modeling for the structural modification of NPs are discussed.

Keywords Natural products, Artificial intelligence, Molecular generative models, Structural modification

[†]Chuan-Su Liu and Bing-Chao Yan have contributed equally to this work.

*Correspondence:

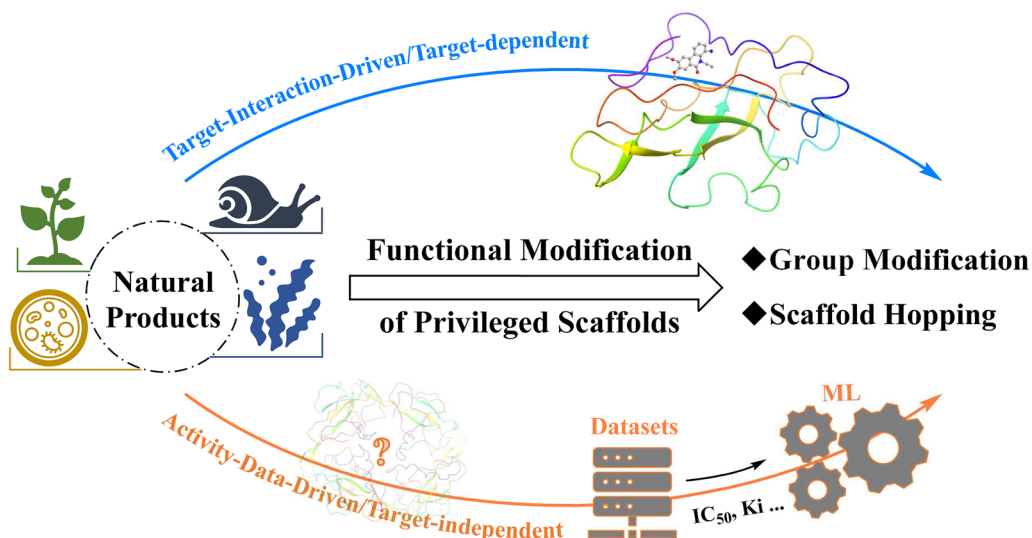
Jin-Cai Lu

jincailu@126.com

Pema-Tenzin Puno

punopematenzin@mail.kib.ac.cn

Graphical Abstract



1 Introduction

Natural products (NPs) have long been regarded as invaluable resources in drug development, consistently yielding innovative leads and drugs for the treatment of human diseases over the past century [1, 2]. Approximately 30% of FDA-approved drugs from 1981 to 2019 originated from natural products (NPs) or their derivatives (NPDs) [3], particularly in the areas of anti-infectives (e.g., penicillin and vancomycin) and anti-tumors (e.g., paclitaxel and camptothecin) [4, 5]. These secondary metabolites derived from plants [6], animals [7], and microorganisms [8] provide invaluable clues for drug discovery due to their unique chemical scaffolds and evolutionarily optimized bioactivities [9]. Despite their remarkable potential in drug development, the clinical applications of NPs face multiple challenges. Their complex stereochemical structures result in unfavorable ADMET properties and violate Lipinski's rule, which often hinders drug development owing to low intestinal absorption and poor oral bioavailability. In addition, most natural products exhibit certain limitations in terms of biological activity, including low potency, limited specificity, and high toxicity, necessitating structural optimization to enhance the efficacy and selectivity [10–12]. Consequently, overcoming the inherent defects of natural products through structural modification to achieve druggability optimization has become a critical challenge in the field of medicinal chemistry.

Structural modification of NPs predominantly focuses on their core scaffolds. The most commonly used

methods include group modification [13], scaffold hopping [14], and structure simplification [15]. Nevertheless, obtaining optimal NPDs continues to pose a significant challenge, even after multiple iterations of structural modifications [10]. Confronted with the significant expenses and inefficiencies of traditional trial-and-error methodologies, as well as the inadequacies of rational design in conventional approaches, computer-aided drug design (CADD) technologies, such as molecular docking, molecular dynamics simulations, and quantitative structure–activity relationship (QSAR) models, have emerged as pivotal tools. These technologies enable the swift evaluation of the affinity of NPDs for target proteins or potential bioactivity, thereby offering more useful guidance for NPs' structural modification [16, 17]. In recent years, with the exponential growth of database scale and breakthroughs in artificial intelligence (AI) algorithms, AIDD has emerged as an evolution of CADD. By fusing generative deep learning with multimodal data, AIDD has realized the fusion of multi-omics data, expansion of generative chemical space, and optimization of dynamic efficacy prediction [18–20]. These advancements are redefining the paradigm of NPs structural modification, transitioning from a “trial-and-error optimization” approach to a “data-driven rational design” strategy. This evolution presents a groundbreaking avenue for addressing the efficiency constraints associated with traditional methods.

Within the AIDD technological framework, molecular generative models, as pivotal methodological

advancements, can be primarily categorized into de novo generation and lead optimization. The latter emphasizes directional structural modifications while preserving the core scaffold [21], which methodologically aligns with the central strategy of NPs structural modification “functional modification of privileged scaffolds” [12, 22]. Building on this methodological synergy, this review systematically categorizes molecular generative models developed from 2020 to 2024, focusing on open-source ones. By integrating case studies of their application in the structural modification of bioactive compounds, we critically review advances in leveraging generative models for NPs optimization. In addition, we examine the existing challenges and discuss potential future pathways for interdisciplinary research.

2 Classification of molecular generation models

The strategy of “functional modification of privileged scaffolds” encompasses two primary approaches for molecular optimization: group modification and scaffold hopping, which operate at local and global levels, respectively. Group modifications, such as side-chain decoration [12] and fragment replacement [23], are directed at specific local regions of molecules (e.g., the modification site). Side-chain decoration focuses on the alteration of acyclic small groups, whereas fragment replacement emphasizes the substitution of functional fragments. Scaffold hopping [24] is primarily aimed at reconstructing the core scaffold. When the core structure of a molecule is redefined as a connected region, both linker design [25, 26] and scaffold hopping functionally and logically manifest as the directed optimization of the molecule’s central connected portion [21].

In the context of the two primary scenarios of “target-known” and “target-unknown” in NPs’ structural modification, the models discussed in the review are further subdivided into “target-interaction-driven” and “molecular activity-data-driven” approaches. The former is predominantly applicable to the structural modification of NPs with known target proteins, thereby enhancing the specificity and success rate of drug development [27, 28]. The latter is applicable not only to scenarios where the disease target protein is unknown [29], but also to the discovery of lead compounds and physicochemical optimization of NPs, offering innovative opportunities for the research and development of NPs.

2.1 Models for functional group modification

These models focus on the structural modification of molecular characteristic regions (e.g., side chains and functional groups) to enhance biological activity (e.g., enhancing target interaction) and improve

physicochemical properties (e.g., solubility and stability) through “fine-tuning” [12].

2.1.1 Target-interaction-driven strategy

This type of models is capable of utilizing protein–ligand interaction data to mine patterns, thereby providing strong support for NP structural modification with known targets [27]. The breakdown for such models focuses on how the generation process interacts and aligns with the target to generate molecules that meet the binding requirements of the target [30] (Table 1).

2.1.1.1 Fragment splicing methods These models select fragments from a predefined chemical fragment library (e.g., pharmacophores and functional groups) and splice them onto the scaffold to generate molecules, ensuring the chemical authenticity and synthesizability of the generated molecules [53, 54]. DeepFrag addresses the challenge of molecule generation by transforming it into a classification task. This is achieved by removing a ligand fragment from a protein–ligand complex and querying the machine learning model to determine the appropriate fragment to be inserted in its place. This allows the model to generate the molecule while considering the surrounding receptor pocket and the full ligand molecule [31] (Fig. 1). Building on fragment generation, FREED combines reinforcement learning (RL) and prioritized experience replay (PER) technology to effectively explore the chemical space and generate pharmacochemically acceptable molecules with high docking scores [32]. The subsequent development of FREED++ fixed and improved the original FREED, resolving multiple implementation errors [33]. DEVELOP combines a graph neural network (GNN) and convolutional neural network (CNN) to construct molecular maps bond-by-bond, and partially constrains the dynamics of molecule generation by utilizing 3D pharmacophore information [34]. STRIFE adopts an architecture similar to that of DEVELOP. It dynamically guides the starting molecule to expand pharmacophore fragments complementary to the target by extracting fragment hotspot maps (FHMs) from the protein target [35]. POEM uses a computer vision method to align protein pockets with PDB-derived images for fragment filtering and linking via DeLinker [36]. DrugGPS learns subpocket prototypes and constructs a global interaction graph to guide the selection of the most suitable fragments from the motif library for addition to the generated molecules [37]. TACOGFN incorporates target pocket information into a generative flow network (GFlowNet) and uses a graph transformer to predict docking scores, generating molecules by gradually adding predefined chemical fragments (72 types) and setting connection points [38]. PGMG uses pharmacophores containing chemical features and spatial distribu-

Table 1 Classification of target-molecule interaction driven models

Models	Architecture	Splicing	Growth	Editing	Notes
DeepFrag [31]*	3D DCNN	✓			Generation to Classification
FREED [32]*	GCN + RL	✓			SAC + PER
FREED++ [33]*	GCN + RL	✓			Simplified & Optimized FREED
DEVELOP [34]*	GNN + CNN	✓			3D Pharmacophore Characterization
STRIFE [35]*	GNN + CNN	✓			FHMs & Dynamic-Extended
POEM [36]*	GGNN	✓			Fragment Alignment & Positioning
DrugGPS [37]*	3D Graph XFMR	✓			Subpocket Interaction
TACOGFN [38]*	GFlowNet	✓			Graph Transformer
PGMG [39]*	GGCN	✓			Predefined Fragments
FRAME [40]*	SE(3)-ENN	✓			Explicit Interaction Modeling
D3FG [41]*	Diffusion + SE(3)-EGN	✓			Rigid Functional Group
AUTODIFF [42]	Diffusion + SE(3)-ECN	✓			Conformal Motif Strategy
MolEdit3D [43]	3D Graph	✓		✓	3D Graph Editing Model
3D-MolGNN _{RL} [44]	G-SchNet + RL		✓		Critic & Dynamic-Feedback
DiffDec [45]*	Diffusion + E(3)-EGNN		✓		Fake Atom Mechanism
AutoFragDiff [46]*	ARDM		✓		Dynamic Fragment Generation
PMDM [47]*	Diffusion + SchNet		✓		Local & Global Diffusion Strategy
DeepICL [48]*	VAE + E(3)-IIN		✓		E(3)-Invariant Interaction Network
DiffInt [49]*	E(3)-Conditional Diffusion		✓		Hydrogen-Bonded Particles
TargetSA [50]*	SA Algorithm		✓	✓	Adaptive Simulated Annealing
Delete [51]*	Unified Masking		✓	✓	Local Editing Model
DrugHIVE [52]*	HVAE		✓		Prior-Posterior Sampling

The symbol [*] indicates that the model is an open-source model

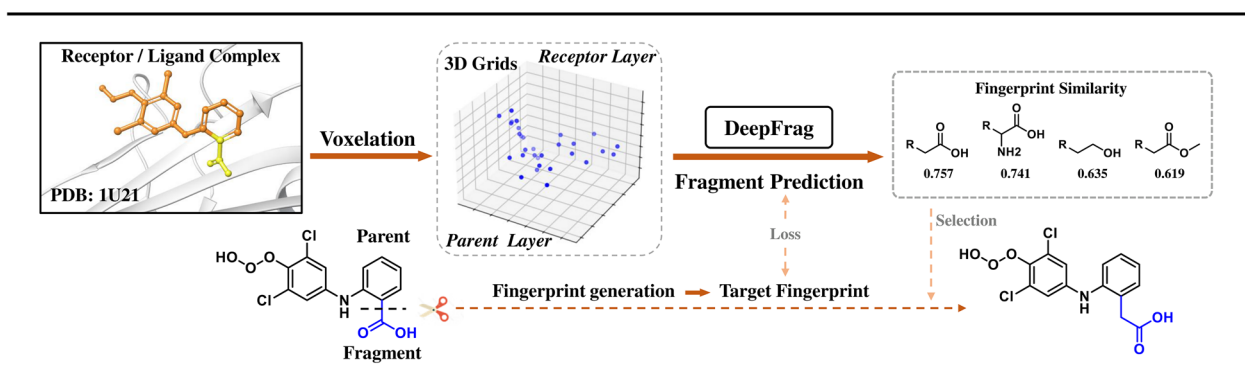


Fig. 1 DeepFrag workflow: The “Parent” and “Fragment” of the initial ligand are highlighted in orange and yellow respectively, which are converted into 3D voxel grids (density channel), merged and fed into the DeepFrag model to predict the missing fragment fingerprints. Finally, the generated fragment fingerprints were compared with a predefined fingerprint library to obtain the prediction results.

tion as the core template, and introduces latent variables to model many-to-many mapping relationships between pharmacophores and molecules to enhance diversity while maintaining target suitability [39]. FRAME utilizes SE(3)-equivariant neural networks to capture the 3D information of target pockets, explicitly model protein–ligand interactions (e.g., hydrogen bonds and π - π stacking), and dynamically select the optimal connection

points and fragments from the starting molecule [40]. D3FG defines molecular fragments as rigid functional groups to preserve the structure of complex fragments, predicts connected fragments by capturing spatial relationships and interactions between proteins and ligands via GNN and uses diffusion modeling to determine fragment position and orientation [41]. AUTODIFF proposes a novel conformal motif strategy to preserve the confor-

mational information of local structures by adding redundant virtual atoms at molecular bond breakage sites [42]. MolEdit3D is a 3D graph-editing model that constructs molecules by stepwise addition or deletion of rigid fragments within the target binding site [43]. These fragment splicing methods, based on different frameworks, utilize protein–ligand interaction information to varying degrees and have successfully completed structure-derived tasks. However, the generated molecules are often limited by the fragment database and cannot fully explore the chemical space of the target binding pocket.

2.1.1.2 Molecular growth methods These models generate molecules directly in the 3D space of the target pocket, maximizing the exploration of the pocket's spatial and target interactions through atom-by-atom or substructure autoregressive generation or global generation based on a diffusion model. 3D-MolGNN_{RL} combines RL with a 3D-Scaffold generative model. Starting from specific 3D scaffolds, the model gradually assesses the binding probability and affinity of generated molecules with the target protein, thus dynamically adjusting the generation process [44]. DiffDec directly models the 3D interactions of molecules with the protein pocket, ensuring that the position and orientation of the generated R motifs are complementary to the pocket shape. The model employs a fake atom mechanism to achieve end-to-end flexible generation of R groups of different sizes within the diffusion model [45]. AutoFragDiff predicts the atomic type and spatial coordinates of new molecular fragments, thus dynamically generating fragments to enhance the local geometric accuracy and binding affinity of the generated molecules [46]. PMDM combines local (simulating covalent bonds) and global molecular dynamics (simulating van der Waals forces), as well as a dual diffusion strategy, to generate 3D molecules that fit a specific target pocket [47]. DeepICL utilizes prior knowledge of protein–ligand interactions to achieve interaction-guided atom-by-atom generation of molecules within the binding pocket [48]. DiffInt introduces interaction particles to explicitly handle hydrogen bond interactions, and uses an E(3)-conditional diffusion model to generate ligand molecules forming hydrogen bonds with the target pocket [49]. TargetSA employs an adaptive simulated annealing (SA) algorithm combined with multi-constraint optimization objectives (e.g., binding affinity, drug-like properties, and synthesizability) to iteratively modify the molecular structure through dynamic local editing (insertion, replacement, deletion, and cyclization) to globally search for the optimal result in a discrete chemical space [50]. Delete (Deep lead optimization enveloped in protein pocket) employs a unified deletion (masking) strategy to dynamically mask atoms or molecular fragments, and then combines them

with the geometric features of protein pockets for repair generation [51]. DrugHIVE uses a hierarchical variational autoencoder (HVAE), combining molecular and protein density grid information to encode multi-scale spatial features (e.g., atom type and hydrogen bond donor/acceptor) for modeling the protein pocket, and achieves lead optimization tasks through spatial prior-posterior sampling [52]. These models explore the chemical space within the protein target pocket using different strategies developed without relying on existing fragment libraries. However, the synthetic accessibility of the generated molecules must be verified through wet experiments.

2.1.2 Activity-data-driven strategy

Examining the structural and activity features of these molecules, which are derived from a recognized dataset of active molecules, enables a thorough investigation of the relationship between the activity and structure [29]. The model predicts the potential activity of NPs by analyzing the common features of active molecules, thereby guiding the structural modification of NPs. The flexibility of the model also enables the exploration of a wider range of chemical spaces and biological activities. For the subcategorization of these models, more attention is paid to how data are input and processed in the model, and how different types of data inputs affect the learning and generative effects of the model (Table 2).

2.1.2.1 SMILES sequence-based methods These models leverage the robust representational capabilities of natural language models (e.g., transformer), facilitating the rapid generation of molecules and demonstrating their suitability for large-scale virtual screening and other application scenarios. However, this model faces illegal issues when multiple side chains are added to the scaffold. To address this problem, the Scaffold Decorator effectively avoids illegal linkages by defining special connection point markers [*] in the SMILES string of the starting scaffold. During the generation phase, the language model grows fragments at these markers in SMILES format, effectively avoiding illegal connections and ensuring that the generated molecular structures are legal and comply with chemical rules. The model provides a data augmentation method that preprocesses all molecules in a small dataset by cutting acyclic bonds (or bonds that comply with RECAP rules), generating a large number of scaffold-decoration tuples and significantly expanding the training data [55]. LibINVENT adopts the same molecular generation strategy as Scaffold Decorator. The model uses a reaction-based preprocessing fragmentation algorithm and combines reinforcement learning (QSAR, ROCS) and a prior model to ensure that the generated compounds adhere to the defined chemical reaction and maintain diversity [56].

Table 2 Classification of molecular activity-data-driven models

Models	Architecture	SMILES	Graph	3D	Notes
Scaffold Decorator [55]*	RNN	✓			Data Augmentation
LibInVENT [56]*	RNN + RL	✓			Reaction-Based Slicing
SAMOA [57]*	RNN + RL	✓			Flexible Sampling
REINVENT 4 [58]*	RNN + TL + RL + CL	✓			Lead Optimization Platform
SAFE [59]*	GPT2-like XFMR	✓			A New Linear Representation
MolGPT [60]*	GPT XFMR	✓			First XFMR Decoder Architecture
EMPIRE [61]*	VAE	✓			Arbitrary Chemical Space
Sc2Mol [62]*	VAE + XFMR	✓			Two-Step Generation Strategy
GraphScaffold [63]*	VAE + GNN		✓		Multi-Property Control Generation
DrugEx v3 [64]*	Graph XFMR + RL		✓		Multi-Objective Constrained
Tree-Invent [65]*	GNN + RL		✓		Topological Tree Constrained
MoLeR [66]*	VAE + GNN		✓		Motif-Based Generation
MRGVAE [67]*	VAE + MPNNs		✓		Molecular Hierarchization
3D-Scaffold [68]*	G-SchNet			✓	Directly Output 3D Coordinates
3D-SMGE [69]*	G-SchNet			✓	ADMET Optimization

The symbol [*] indicates that the model is an open-source model

SAMOA modifies the recurrent neural network (RNN) sampling process, directly enforcing molecules to follow a predefined scaffold during generation, and uses reinforcement learning (QSAR) to optimize the pharmacological activity and pharmacokinetic properties of the generated molecules [57]. REINVENT 4 trains pre-trained models on large public datasets such as ChEMBL and PubChem. Transfer learning (TL) facilitates the concentration on tasks involving small datasets, while reinforcement learning (RL) and course learning (CL) empower the model to generate molecules with specific properties, thereby progressively enhancing the efficiency and quality of molecular generation [58]. SMILES-based models benefit from their concise text format, which facilitates the storage and transmission of molecular structure data but are also limited by their strict syntactic rules. SAFE proposes a new molecular linear representation method that represents SMILES strings as unordered sequences of interconnected fragment blocks, effectively bypassing complex decoding schemes and simplifying the generation task. The SAFE-GPT model was trained on a dataset containing 1.1 billion SAFE representations and can perform target-oriented lead compound optimization [59].

2.1.2.2 Molecular graph-based methods These models represent molecules as graph structures with nodes as atoms and edges as chemical bonds, capturing interatomic bonding relationships directly and avoiding the limitations of the SMILES syntax. GraphScaffold starts from a given molecular scaffold and extends it by sequentially adding atoms and bonds. The model can control molecular properties through a conditional generation process

[63]. DrugEx v3 combines a graph transformer and RL to process molecular structure information more effectively. The model can deal with more complex scaffold structures and simultaneously optimize multiple objectives (e.g., QED and affinity) [64]. Tree-Invent combines molecular graph topological trees with RL. Model uses three independent sub-models to perform Nodeadd, Ringgen, and Nodeconn operations. This approach achieves precise control over the molecular generation process and optimizes the results according to target properties [65].

2.1.2.3 3D structure-based methods These models combine atomic coordinates and chemical bonding information to generate molecules with spatial conformations, effectively addressing the lack of conformational changes in representations based on SMILES and molecular graphs. 3D-Scaffold uses a feature learning module to extract rotation and translation-invariant atomic features, capturing the incomplete chemical environment. The atomic placement module then predicts the type of the next atom (discrete distribution) and its 3D distance distribution with all placed atoms, considering the dynamic interactions in molecular generation and avoiding the molecular rigidity assumption [68]. This type of models relies on high-quality 3D structural data and requires to process multi-modal data involving continuous spatial coordinates and discrete atomic types, resulting in extremely high computational complexity.

From the performance verification experiments of these models, it is evident that activity-data-driven molecular generation models exhibit a certain degree of activity-oriented nature. For instance, Scaffold

Decorator, when applied to the DRD2 active molecule set (4,211 molecules), generated molecules from different scaffolds that had a significantly higher proportion of predicted activity (based on a pre-trained activity prediction model) compared to random decoration using the ChEMBL dataset [55]. SAMOA on the MMP-12 inhibitor lead optimization dataset, 23% of the molecules generated by the model met the constrained scaffold and high activity (pre-trained QSAR predicts $pIC_{50} > 7.5$) [57]. For REINVENT 4, in the task of generating new PDK1 inhibitors, the proportion of active molecules generated through RL was 1.9% (with a docking score ≤ -8 kcal/mol and QED ≥ 0.7). After using TL-RL, this proportion increased to 3.5% (the proportion of active molecules based on RL and RL-TL increased with the number of model runs) [58]. DrugEx v3 carried out targeted molecular generation for A2AAR, with 86.1% of the generated molecules predicted to be active [64]. 3D-Scaffold targeted NSP15 and generated multiple molecules based on Exebryl-1, which showed good binding affinity in docking simulations and had lower SA scores and higher QED scores [68].

2.1.2.4 Other methods The following models focus on scaffold constraints and the optimization of physicochemical properties. These models are trained on publicly available large datasets to enhance the diversity of molecular generation results and explore a wider chemical space, making them more suitable for lead compound discovery.

SMILES-based models. MolGPT learns the SMILES representation and structural features in molecular datasets (MOSES and GuacaMol) through a masked self-attention mechanism, enabling it to perform conditional generation and control molecular properties (e.g., Log P, TPSA, SAS, and QED) [60] (Fig. 2A). EMPIRE generates new fragments through VAE and a building block list, resulting in novel molecules in arbitrary chemical spaces.

For example, EMPIRE can generate molecules containing unique scaffolds (e.g., bicyclo[1.1.1]pentane and cubane) or elements (e.g., boron and silicon), which facilitates virtual screening in unexplored chemical spaces and enhances the efficiency of drug design [61] (Fig. 2B). Sc2Mol decomposes molecular generation into two steps: scaffold generation and scaffold decoration, which are handled by the VAE and transformer, respectively. The model learns the structural features of molecules from molecular datasets (MOSES and ZINC-250 k) and supports the generation of molecules with specific properties from a given scaffold [62].

Molecular graph-based models. MoLeR adds atoms or predefined motifs to a complete scaffold with single bonds under strict scaffold-constrained conditions. The model combines the molecular group optimization (MSO) method and can generate molecules with target properties [66]. MRGVAE decomposes molecules in the dataset (ChEMBL) into small fragments and groups them into interchangeable fragment clusters based on the local structural environment of the fragments. The model combines random sampling and top-p sampling to increase fragment diversity while controlling the distribution of chemical properties. The model controls the generation of multiple molecular attributes, such as scaffold structure, molecular weight, and log P, through conditional generation [67].

3D structure-based models. 3D-SMGE generates 3D molecular structures from a specific scaffold and evaluates their ADMET properties by learning molecular structures and properties from a large molecular dataset [69].

2.2 Models for scaffold hopping

Local group modification can enhance functional adaptability, but is limited by the inherent defects of the original scaffold (e.g., non-modifiable metabolic sites). Therefore, scaffold hopping is required to achieve a leap in drug-like

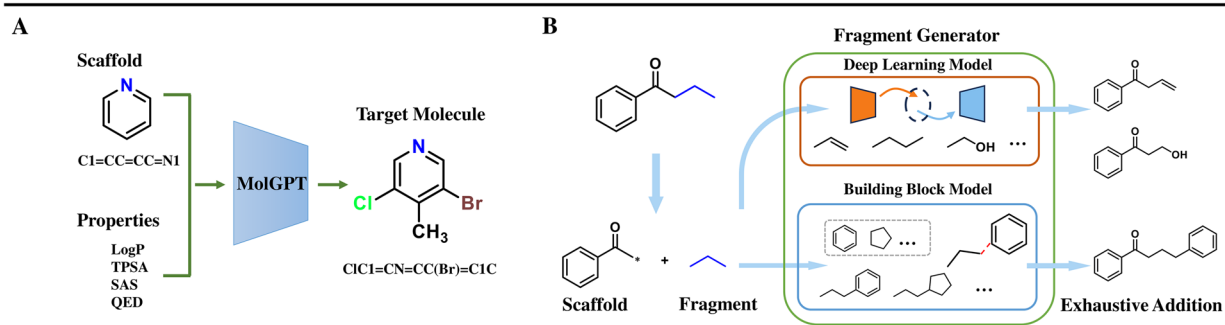


Fig. 2 Two representative data-driven methods. **A** MolGPT illustration and **B** EMPIRE illustration

properties. These models mainly involve global adjustments to the core scaffold or connecting parts of active compounds to optimize their structure and function, thereby breaking through the activity ceiling of the original scaffold [15] (Table 3).

2.2.1 Target-interaction-driven strategy

Focusing on scaffold-hopping molecular generation models, DeepHop integrates the sequence information of the target protein and the 3D conformation information of molecules. This allows the model to generate molecules with similar 3D structures but different 2D scaffolds while considering the target protein information. However, the side chains that need to be retained may be modified in the process [70] (Fig. 3A). To address this limitation, GraphGMVAE inputs the bioactivity condition vector of the protein target together with the scaffold and side-chain embeddings of the molecule into the generative model; however, the GraphGMVAE model is not open-source, which limits its widespread application [71]. The open-source model ScaffoldGVAE, trained on the ChEMBL database and focusing on specific kinase protein data, maps scaffold embeddings to a Gaussian mixture distribution while keeping side-chain

embeddings unchanged. This enables the generation and optimization of scaffolds [72]. DiffHopp guides scaffold hopping in a given protein pocket by learning the interaction information of a large number of protein–ligand complexes in PDBbind [73]. DiffLinker is a target-driven linker design model that uses the atomic point clouds of the protein pocket as fixed information and inputs them into the neural network along with the atomic point clouds of the molecular fragment. As a result, the generated molecules can consider the geometric constraints of the protein pocket, thus generating molecules compatible with the pocket structure [74].

2.2.2 Activity-data-driven strategy

All the models in this category are linker design models that also take into account scaffold hopping. As the first linker design model to incorporate 3D structural information into the generation process, DeLinker learns the molecular structure and property information from the ZINC and CASF datasets. The model generates molecules by utilizing the relative distance and orientation information between two fragments or partial structures [75]. Benefiting from the expressive power of transformer, SyntaLinker transforms the fragment linking task

Table 3 Classification of scaffold hopping methods

Methods	Architecture	Scaffold Hopping	Linker Design	Notes
Target-interaction-driven				
DeepHop [70]*	3D GNN + XFMR	✓		3D Structural Similarity
GraphGMVAE [71]	GMVAE + DualMPNN	✓		Side-Chain Retention
ScaffoldGVAE [72]*	VAE + MVGNN	✓		Interaction Constraints
DiffHopp [73]*	E(3)-E Graph Diffusion	✓		PDBbind Data Learning
DiffLinker [74]*	E(3)-E Diffusion	✓	✓	Geometric Constraints
Activity-data-driven				
DeLinker [75]*	GGNN	✓	✓	Distance & Direction
SyntaLinker [76]*	XFMR	✓	✓	NLP Task
SyntaLinker-Hybrid[77]*	XFMR + TL	✓	✓	Fragments Hybridization
DRLinker [78]*	XFMR + RL	✓	✓	RL-Based Scoring
GRELInker [79]*	GGNN + RL + CL	✓	✓	Outperforms DRLinker
Link-INVENT [80]*	RNN + RL	✓	✓	Fragment Splicing

The symbol [*] indicates that the model is an open-source model

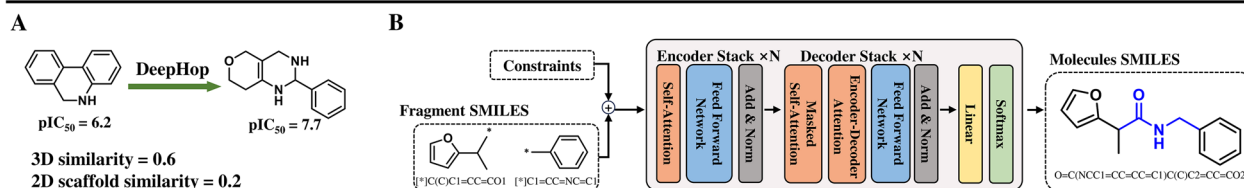


Fig. 3 Two representative scaffold hopping methods. **A** DeepHop illustration and **B** SyntaLinker illustration

into a NLP-like task, by parsing the syntactic patterns and fragment linking rules of molecular SMILES sequences in the database and introducing a multi-conditional control mechanism (e.g., the shortest linking distance, the presence or absence of hydrogen bond donors, acceptors, rotatable bonds, and loops). This allows for end-to-end controllable generation from a given fragment pair to a complete molecule [76] (Fig. 3B). Subsequently, SyntaLinker-Hybrid was developed based on the generative model of SyntaLinker, which uses TL to fine-tune target-focused active compound data and combines fragment hybridization technology to generate molecules with target specificity [77]. DRlinker combines transformer and RL to generate compounds under the guidance of a given scoring function and can control specific properties of the generated compounds, such as linker length, log P value, bioactivity, QED, SAS, 3D similarity, and 2D structural novelty [78]. Compared with DRlinker, GRELinker combines the gated graph neural network (GGNN) architecture with RL and CL algorithms. The model has increased the percentage of molecules that meet property constraints and is able to generate more complex linker structures [79]. Link-INVENT is an extension of the existing de novo molecular design platform REINVENT [81] and is specialized for tasks such as fragment linking, scaffold hopping, and PROTAC design. The model is trained on a large number of molecular datasets (e.g., the ChEMBL database) and controls multiple attributes of generated molecules, such as scaffold structure, molecular weight, and log P, through conditional generation [80].

3 Application cases of structural modification

3.1 Group modification cases

Local group modification models optimize target affinity by fine-tuning side chains or functional groups while retaining the core scaffold. The following cases demonstrate their applications in fields such as antiviral and anticancer research.

3.1.1 Cases of target-interaction-driven generation

3.1.1.1 DeepFrag for target optimization in antiviral drug development SARS-CoV-2 N protein is a key target for antiviral drug development [82]. Hao et al. identified the antiviral activity of phenanthridine derivatives through CADD design validation [83] and identified the binding site as the N-terminal domain of the N protein. To obtain lead compounds with higher affinity for N protein and stronger antiviral activity, researchers further employed the DeepFrag model (Sect. 2.1.1) to perform directed replacement of the side chains of phenanthridine alkaloids with antiviral potential based on the local environment of the N protein binding pocket. To balance target-orientation and diversity, the EMPIRE model (Sect. 2.1.2) was

introduced to enhance the generative chemical space. The models generated 16,689 virtual derivatives, and 44 compounds were synthesized after virtual screening and molecular docking. Compound **38** exhibited high affinity and significant antiviral activity in vitro (EC_{50} = 11.3 μ M, $TI > 17.7$) [84] (Fig. 4A).

3.1.1.2 DeepFrag for optimization of anticancer drug Topo II α inhibitors Human DNA topoisomerase II α (topo II α) is an important target for anticancer drugs [85]. In a previous work by Perdihi et al. 4,6-substituted-1,3,5-triazin-2(1H)-ones with strong inhibitory activity against topo II α were discovered through a combination of structure-based and ligand-based pharmacophore models and molecular docking calculations [86]. To enhance their binding affinity and inhibitory activity, researchers used DeepFrag to analyze the ATP-binding site of topo II α and guide the optimization of the R group of the triazine compounds. They also used molecular docking and molecular dynamics (MD) simulations to screen and synthesize 44 derivatives. Compounds **32** and **33** showed a significant increase in inhibitory activity against topo II α (IC_{50} = 8.7 μ M and 7.7 μ M, respectively), and exhibited strong cytotoxicity in glioblastoma and breast cancer cell lines [87] (Fig. 4B).

3.1.2 Cases for activity-data-driven generation

3.1.2.1 Scaffold Decorator for the discovery of selective antagonists of adenosine A_{2B} receptors The adenosine A_{2B} receptor (A_{2B}AR) is associated with inflammatory diseases, but the development of its antagonists often fails due to poor subtype selectivity [88]. Lei et al. used an activity-data-driven Scaffold Decorator model (Sect. 2.1.2), based on a dataset of known active molecules, to diversify derivatizations of the 3-amine and 3-benzyl scaffold. They generated over 90,000 virtual molecules and used virtual screening, ADMET screening, and QSAR prediction to identify the lead compound **ABA-1266**, which was screened for its free energy of binding to A_{2B}AR of -69.41 kcal/mol and showed high selectivity among multiple adenosine receptor subtypes [89] (Fig. 5A).

3.1.2.2 Scaffold Decorator for the discovery of DDR1 selective inhibitors The receptor tyrosine kinase DDR1 plays a key role in cell signaling, tissue development, and disease progression. However, the development of DDR1 inhibitors is often limited by their poor target specificity [90]. Zheng et al. found that **DC-1**, discovered in previous studies, has inhibitory activity against DDR1 (inhibition rate of 48%). Molecular docking analyses have revealed the potential for optimization of the binding model of **DC-1** and DDR1 [91]. To obtain compounds with high activity specificity, researchers have integrated Scaffold Decora-

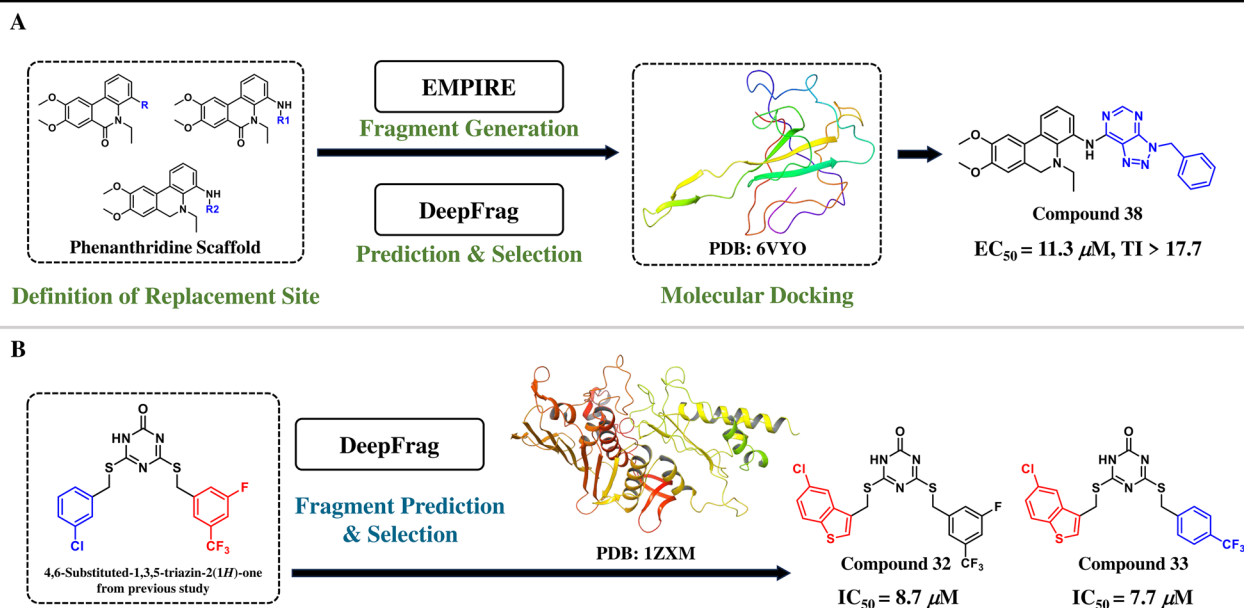


Fig. 4 Two application examples of Deepfrag. **A** Antiviral drug development and **B** Topo II α inhibitors optimization

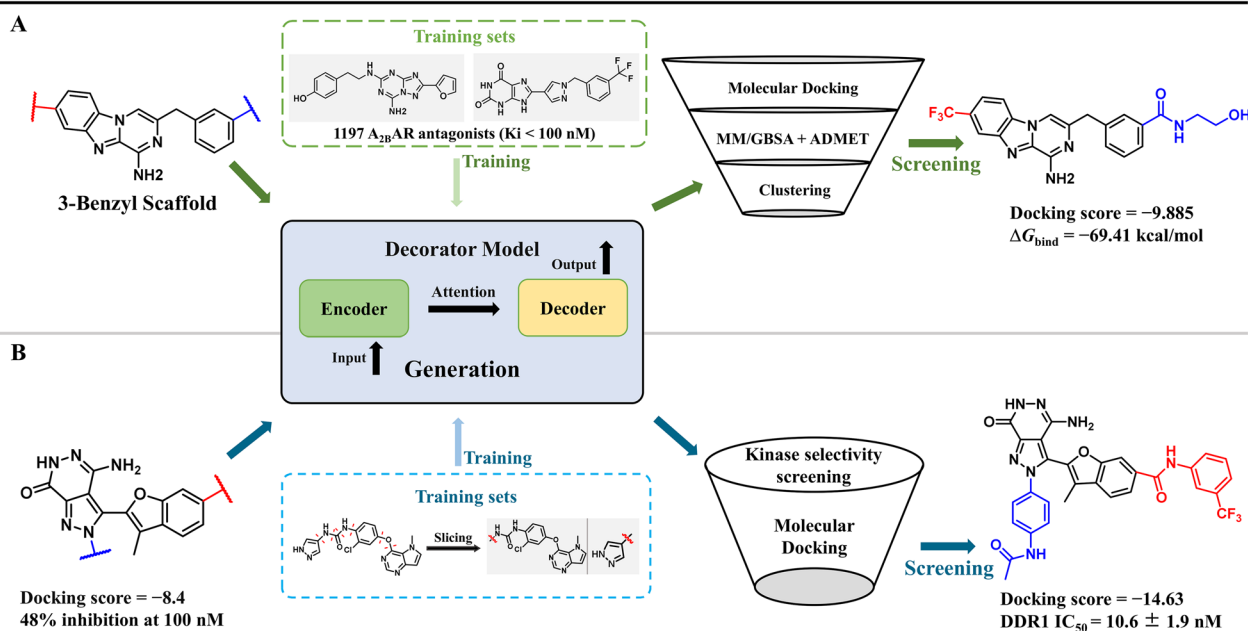


Fig. 5 Two application examples of Scaffold Decorator. **A** Discovery of A_{2B} receptors and **B** DDR1 inhibitors

tor molecular generation, kinase selectivity screening, and molecular docking to identify a novel DDR1 inhibitor, compound 2. The compound exhibited excellent selectivity ($S(10) = 0.002$ out of 430 kinases) with an IC_{50} value of 10.6 ± 1.9 nM for DDR1 and significantly inhibited pro-

inflammatory cytokines and DDR1 autophosphorylation [92] (Fig. 5B).

3.1.2.3 LibINVENT for the discovery of new inhibitors of Cbl-b The E3 ubiquitin ligase Cbl-b plays multiple

roles in immune and tumor cells. Inhibiting its function can enhance the antitumor capacity of the immune system and suppress the growth and survival of tumor cells [93]. To identify new Cbl-b inhibitors, Hughes et al. utilized the LibINVENT model (Sect. 2.1.2) with a triazolone core scaffold and optimized pharmacophore matching (ROCS score) and drug-like properties ($QED > 0.6$) through RL. Through virtual screening and ADMET property prediction, the researchers selected **VC1** and **VC2**, which exhibited good target selectivity. Further structural optimization and in-vitro validation revealed that compounds **20** and **24** had IC_{50} values of 1.4 μM and 0.17 μM , respectively, and demonstrated significant cytotoxicity in various human cancer cell lines [94].

3.1.2.4 SAMOA for dynamic optimization of ATM kinase inhibitors ATM kinase is a key target of the DNA damage repair pathway and its inhibitors have important potential in anticancer therapy [95]. Chen et al. proposed a dual-modal optimization strategy, utilizing the flexible framework of the SAMOA algorithm (Sect. 2.1.2). Based on a RL process with a composite scoring function of target specificity (SBDD) and ligand similarity (LBDD), 20,194 virtual molecules were generated from the 6-(pyridin-3-yl) quinoxaline scaffold. After molecular docking and ADMET prediction, lead compound **7a** was selected, which exhibited an IC_{50} value of 5 nM for ATM inhibition in vitro but had poor metabolic stability. By replacing different urea groups, compound **8d** ($IC_{50} = 2$ nM) exhibited better ATM inhibitory activity and partial metabolic sta-

bility. Further optimization led to compound **10r**, which showed high selectivity with an IC_{50} value of 1 nM for ATM kinase, almost no inhibitory effect on 103 other kinases (showing high selectivity), and excellent metabolic stability and cellular activity in vitro [96].

3.2 Scaffold hopping cases

Scaffold hopping models overcome the limitations of the original scaffold. The following cases demonstrate the application of these methods in the development of kinase inhibitors.

3.2.1 Target-interaction-driven generation for scaffold hopping in JAK1 inhibitors

The non-receptor tyrosine kinase JAK1 plays a key role in cell signaling in immune responses and inflammatory processes [97]. The development of new JAK inhibitors remains highly focused on improving subtype selectivity and enhancing therapeutic effects while reducing toxicity. Huang et al. developed the GraphGMVAE model (Sect. 2.2.1) to generate novel JAK1 inhibitors for validation starting from the FDA-approved upadacitinib. Among the generated molecules, 97.9% had novel scaffolds that differed from those of known JAK1 inhibitors. Seven compounds were synthesized, with compound **Ten01** showing an IC_{50} value of 5.0 nM for JAK1 inhibition [71] (Fig. 6A).

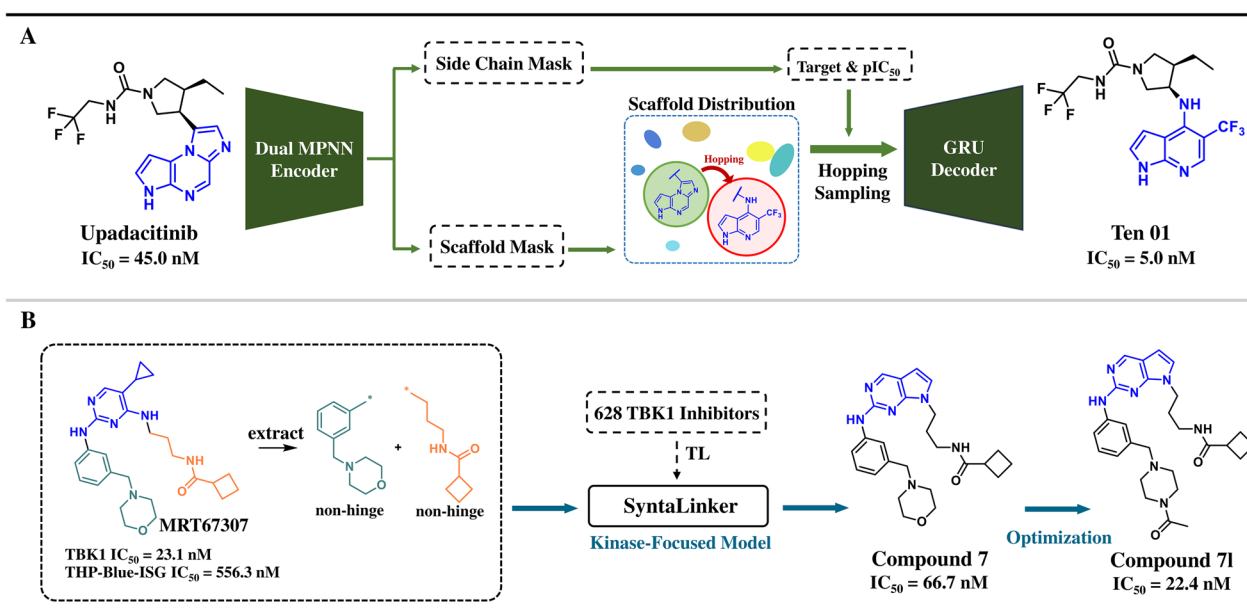


Fig. 6 **A** Scaffold hopping of JAK1 inhibitors and **B** linker optimization of TBK1 inhibitors

3.2.2 Activity-data-driven generation for linker optimization in TBK1 inhibitors

TANK-binding kinase 1 (TBK1) plays an important role in the innate immune system and is involved in the development of several cancers [98]. Although several small-molecule TBK1 inhibitors have been reported, none have been used clinically. Lu et al. first trained a prior model on a general ChEMBL fragment set and then performed TL on a kinase inhibitor dataset to construct a kinase-specific SyntaLinker model (Sect. 2.2.2). Researchers generated a series of molecules by replacing the non-hinge binding fragments of the known TBK1 inhibitor **MRT67307** and selected lead compound **7** through molecular docking. Compound **7** exhibited an IC_{50} value of 66.7 nM for TBK1 inhibition and demonstrated good target selectivity. Further structural optimization and in vitro validation led to the identification of compound **7I**, which exhibited an IC_{50} value of 22.4 nM for TBK1 inhibition and significantly inhibited the expression of TBK1 downstream genes in THP1 cells [99] (Fig. 6B).

3.3 Discussion

The limitations in the application of the model are analyzed in relation to the application cases and the classification of the models in Sect. 2.

Firstly, DeepFrag matches fragments from a predefined fragment library and is unable to generate new fragments. It also assumes that fragment addition will not significantly change the binding geometry of the parent molecule, without considering the dynamic changes between the molecule and the binding pocket. In contrast, the pocket-based molecular growth methods (e.g., DiffDec [45], AutoFragDiff [46], and DiffInt [49]) demonstrate the ability to handle interactions with pocket during the molecular generation process, effectively avoiding the limitations of fragment library and the assumption of molecular rigidity.

Secondly, although the activity-data-driven SAMOA model integrates the 3D structural information of ATM kinase through RL, its SBDD module is based on a static crystal conformation and does not simulate the dynamic conformational effects of proteins. Molecular generation models for dynamic interaction modeling (e.g., 3D-MoLGNNRL [44], FRAME [40], and AutoDiff [42]) achieve dynamic control of the molecular generation process based on different architectures, overcoming the limitations of static target modeling and the absence of dynamic interactions.

Thirdly, the above cases of molecular generation methods heavily rely on CADD virtual screening for activity validation, resulting in a disconnect between the generation and validation stages and failing to achieve end-to-end dynamic optimization. For example, Scaffold

Decorator performs the task of scaffold decoration and performs a certain activity-driven molecule generation based on preprocessed activity datasets, but requires validation screening by molecular docking. The same activity-data-driven lead optimization model, REINVENT 4 [58], which combines TL, RL, and CL, has improved the efficiency and quality of molecule generation.

Finally, models such as EMPIRE and Scaffold Decorator lack explicit modeling of synthetic pathways (such as reaction conditions and yields). Insufficient feasibility prediction leads to the fact that the generated molecules often cannot enter the experimental validation stage because of high synthetic difficulty. Models that incorporate chemical reaction rules into the preprocessing process, such as LibINVENT [56], constrain the synthetic feasibility of generated molecules through predefined reaction types, thereby improving the synthetic feasibility of molecules.

4 Conclusion and prospects

In response to the scarcity of application cases of molecular generation models in the field of NPs, this paper integrates limited research cases and combines successful practices in the field of synthetic drug molecules to verify their cross-domain applicability and technical feasibility. For example, Hao et al. successfully obtained compounds with high affinity and significant antiviral activity by targeted optimization of the side chains of phenanthridine alkaloids based on the DeepFrag model, confirming the efficiency of AIDD in the targeted modification of NPDs [84]. Further analysis shows that molecular generation models have multi-dimensional technical potential in optimizing the structures of NPs: **a) Efficient exploration of chemical space:** for example, Scaffold Decorator generates more than 90,000 virtual molecules from predefined modification sites of the scaffolds, significantly improving the efficiency of lead compound discovery [89]; **b) Design of novel scaffolds:** for example, GraphGMVAE generates novel molecules from the known JAK1 inhibitor upadacitinib, 97.9% of which contain novel scaffolds, breaking through patent limitations and providing a new paradigm for NP-based scaffold innovation [71]; **c) Target-driven precise optimization:** for example, SAMOA combines the 3D structure of ATM kinase with ligand similarity to generate a highly selective inhibitor compound **10r** ($IC_{50}=1$ nM), which shows no significant off-target effects on other 103 kinases [96]; **d) Activity-oriented molecular generation:** for example, SyntaLinker integrates the TBK1 kinase inhibitor dataset through TL to generate the target-selective inhibitor **7I**, which has an IC_{50} value of 22.4 nM for TBK1 inhibition. It offers a reference for applying

activity-data-driven models to NPs with unknown mechanisms of action [99]. These cases demonstrate the feasibility and applicability of molecular generation models for optimizing the structures of natural products. Although progress has been made, contemporary research continues to depend significantly on validation experiences derived from synthetic drug development. Therefore, there is an urgent need to develop NP-specific generation strategies and evaluation systems.

The application of target-interaction-driven models in NPs is limited by the following: **a) High data acquisition costs:** large amounts of high-quality protein–ligand complex data (e.g., SPR and X-ray crystallography) or from limited public databases (e.g., BindingDB) are required. However, NP-protein complex structures are extremely scarce and expensive to validate experimentally [100]; **b) Excessive target dependence:** the predictive ability of a model is highly dependent on the structural and functional information of known targets. If the conformational changes in the target protein (e.g., metastable effects) are not adequately characterized, or if there are multi-target synergistic effects, the model may fail to accurately predict the binding affinity; **c) Limited generalization:** the model generalizes poorly to new or cross-species targets; **d) Poor dynamic adaptation:** target proteins may change their binding properties owing to the cellular microenvironment (e.g., pH and ion concentration) or post-translational modifications (e.g., phosphorylation). Static datasets are difficult to cover these dynamic changes, leading to prediction deviations [101].

The application limitations of activity-data-driven models in NPs include the following: **a) Data bias and noise sensitivity:** the over-representation of positive molecules and the lack of negative results in the activity dataset lead to model overfitting of the known chemical space and limited diversity of generated molecules. Meanwhile, the differences in the experimental conditions (e.g., pH and temperature) of the activity data reported in different studies are not standardized, which affects the model generalization performance of the prediction model. **b) Lack of mechanistic interpretability:** the model correlates structure and activity through statistical regularities but cannot to clarify specific molecular mechanisms of action (e.g., target binding mode and signaling pathway regulation), limiting its application in mechanism-oriented optimization. **c) Insufficient chemical space coverage:** existing activity datasets mainly focus on specific compound categories (e.g., small molecules), with insufficient characterization of complex natural products (e.g., macrolides and polyketides) and novel synthetic molecules, leading to undesirable model performance in the prediction of new chemical entities [102, 103].

Currently, molecular generation models face several common challenges in NPs applications, including: **a) High computational resource dependency:** a large number of computational resources are required for training, especially for deep learning models (e.g., transformer-based models), which restricts their application in resource-limited scenarios; **b) Inadequate modeling of complex systems:** neither target-interaction nor activity-data-driven models can fully simulate the complexity of real biological systems (e.g., gene interaction networks and metabolic pathway regulation), leading to discrepancies between predictions and in-vivo effects [104]; **c) Explainability and ethical risks:** models generate molecules through implicit associations, and their structure–activity relationships lack transparent resolution. They may produce results that do not align with biological common sense or ethical guidelines (e.g., toxic molecules), necessitating expert knowledge validation.

The intelligent transformation of NPs structural modification faces a historical opportunity, but requires systematic breakthroughs in data, algorithms, and technology. NPs-based molecular generation is characterized by small sample data and most of the targets have not been elucidated. **a) Sample learning and data augmentation:** for example, constructing virtual molecular libraries based on reaction rule-driven molecular modularization [105] or fragmentation [55, 56] to generate chemically reasonable new molecules from limited data; Also, using pre-trained models trained on general molecular libraries (like ZINC and ChEMBL), and fine-tuning them for NPs datasets through TL [58, 77], and combining with an active learning strategy [106], to prioritize the labeling of high-potential molecules and significantly reduce the cost of experimental validation. **b) Dynamic interaction modeling and multi-modal fusion:** for example, introducing equivariant graph neural network and adaptive molecular dynamics to simulate the effect of target allosteric effects on binding stability [107–109]; Constructing “structure–activity–pathway” multi-level prediction models to integrate target interaction networks, transcriptomics and metabolomics data to analyze the multi-target synergistic mechanism [110]. **c) Lightweight model architecture:** large parametric models are compressed into lightweight versions by knowledge distillation techniques to reduce computational resource consumption while maintaining generation performance [111, 112]. **d) Closed-loop automation systems:** the automated synthesis platform integrates inverse synthesis analysis and automated synthesis platform [113, 114], which has been applied to synthesize compounds (e.g. small molecule drugs). In the future, to solve the structural complexity and build NPs-specific synthetic database

for algorithm optimization, deep learning molecular design, automated synthesis, biosynthesis pathway, modular reactor, and high-throughput activity detection should be integrated to build a full-process system of “virtual design → robotic synthesis → experimental feedback” and realize the closed-loop of generation-synthesis-verification of NPs synthesis and structural modification. **e) NPs Databases:** existing NPs databases (for example, NPASS, SuperNatural 3.0, LOTUS, and COCONUT) have formed a complete system in chemical structure analysis and physicochemical property characterization. However, they have limitations such as single-dimensional biological activity data, missing dynamic synthetic pathways, and annotation limitations. Future databases need to be upgraded to intelligent predictive platforms by integrating multi-modal data and enabling real-time updates (blockchain tracking) and embedding AI toolchains to transform repositories into intelligent synthetic activity prediction centers.

AIDD is promoting the structural modification of NPs from “trial-and-error optimization” to “data-driven rational design”. Future efforts in lightweight architectures, multi-modal fusion, dynamic modeling, and closed-loop automation may overcome key challenges, such as data scarcity, low synthetic feasibility, and multi-objective conflicts, and promote the efficient transformation of NPs from “chemical entities” to “clinical drugs”.

Acknowledgements

We acknowledge the researchers in the field of artificial intelligence-driven drug design for their significant contributions to advancing this emerging area of research.

Author contributions

Pema-Tenzin Puno and Jin-Cai Lu conceived and designed the review; Chuan-Su Liu and Bing-Chao Yan collected and analyzed the data and wrote the initial draft of the manuscript; Pema-Tenzin Puno, Jin-Cai Lu, Han-Dong Sun provided critical feedback and helped revise the manuscript; All authors read and approved the final manuscript.

Funding

This work was financially supported by the National Science Fund for Distinguished Young Scholars (82325047), Regional Innovation and Development Joint Fund of NSFC (U24A20807), Youth Innovation Promotion Association CAS (2023411), National Natural Science Foundation of China (22477123), Major Projects for Fundamental Research of Yunnan Province (202201BC070002), CAS “Light of West China” Program and CAS Interdisciplinary Innovation Team (xbzg-zdsys-202303), Yunnan Revitalization Talent Support Program: Yunling Scholar Project, Yunnan Province Science and Technology Department (202305AH340005).

Availability of data and materials

All the data and materials provided in the manuscript are obtained from included references and available upon request.

Declarations

Ethics approval and consent to participate

Ethical declaration is not applicable for this review.

Competing interests

Pema-Tenzin Puno is the journal's editor but was not involved in the peer review or decision making process in this article. The other authors declare no conflicts of interest.

Author details

¹School of Traditional Chinese Materia Medica, Shenyang Pharmaceutical University, Shenyang 110016, Liaoning, China. ²State Key Laboratory of Phytochemistry and Natural Medicines, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, China.

Received: 13 April 2025 Accepted: 13 May 2025

Published online: 06 June 2025

References

- Cragg GM, Newman DJ. Natural products: a continuing source of novel drug leads. *Biochem Biophys Acta*. 2013;1830:3670–95.
- Baker DD, Chu M, Oza U, Rajgarhia V. The value of natural products to future pharmaceutical discovery. *Nat Prod Rep*. 2007;24:1225–44.
- Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod*. 2020;83:770–803.
- Brown DG, Lister T, May-Dracka TL. New natural products as new leads for antibacterial drug discovery. *Bioorg Med Chem Lett*. 2014;24:413–8.
- Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov*. 2015;14:111–29.
- Hai Q-X, Hu K, Chen S-P, Fu Y-Y, Li X-N, Sun H-D, He H-P, Puno P-T. Silvaticusins A–D: *ent*-kaurane diterpenoids and a cyclobutane-containing *ent*-kaurane dimer from *Isodon silvaticus*. *Nat Prod Bioprospect*. 2024;14:1–8.
- Rashad M, Sampò S, Cataldi A, Zara S. Biological activities of gastropods secretions: snail and slug slime. *Nat Prod Bioprospect*. 2023;13:1–9.
- Banerjee S, Cabrera-Barjas G, Tapia J, Fabi JP, Delattre C, Banerjee A. Characterization of Chilean hot spring-origin *Staphylococcus* sp. BSP3 produced exopolysaccharide as biological additive. *Nat Prod Bioprospect*. 2024;14:1–16.
- Eberhardt L, Kumar K, Waldmann H. Exploring and exploiting biologically relevant chemical space. *Curr Drug Targets*. 2011;12:1531–46.
- Chen J-C, Li W-L, Yao H-Q, Xu J-Y. Insights into drug discovery from natural products through structural modification. *Fitoterapia*. 2015;103:231–41.
- Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. *Nat Chem*. 2016;8:531–41.
- Yao H, Liu J-K, Xu S-T, Zhu Z-Y, Xu J-Y. The structural modification of natural products for novel drug discovery. *Expert Opin Drug Discovery*. 2017;12:121–40.
- Haynes RK, Fugmann B, Stetter J, Rieckmann K, Heilmann H-D, Chan H-W, Cheung M-K, Lam W-L, Wong H-N, Croft SL, Vivas L, Rattray L, Stewart L, Peters W, Robinson BL, Edstein MD, Kotecka B, Kyle DE, Beckermann B, Gerisch M, Radtke M, Schmuck G, Steinke W, Wollborn U, Schmeier K, Römer A. Artemisone—a highly active antimalarial drug of the artemisinin class. *Angew Chem Int Ed*. 2006;45:2082–8.
- Pereira AR, Strangman WK, Marion F, Feldberg L, Roll D, Mallon R, Hollander I, Andersen RJ. Synthesis of phosphatidylinositol 3-kinase (PI3K) inhibitory analogues of the sponge meroterpenoid liphagal. *J Med Chem*. 2010;53:8523–33.
- Wang S-Z, Dong G-Q, Sheng C-Q. Structural simplification of natural products. *Chem Rev*. 2019;119:4180–220.
- Gromiha MM, Harini K. Protein-nucleic acid complexes: docking and binding affinity. *Curr Opin Struct Biol*. 2025;90:1–9.
- Panigrahi D, Sahu SK, Panigrahi D, Sahu SK. Computational approaches: atom-based 3D-QSAR, molecular docking, ADME-Tox, MD simulation and DFT to find novel multi-targeted anti-tubercular agents. *BMC Chem*. 2025;19:1–28.
- Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, Volkov Y, Zholus A, Shayakhmetov RR, Zhebrak A, Minaeva LI,

- Zagribelnyy BA, Lee LH, Soll R, Madge D, Xing L, Guo T, Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038–40.
19. Jiang D, Hsieh C-Y, Wu Z-X, Kang Y, Wang J, Wang E, Liao B, Shen C, Xu L, Wu J, Cao D-S, Hou T-J. Interactiongraphnet: a novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *J Med Chem*. 2021;64:18209–32.
 20. Cao D-H, Chen M-A, Zhang R-Z, Wang Z-K, Huang M-L, Yu J, Jiang X-Y, Fan Z-H, Zhang W, Zhou H, Li X-T, Fu Z-Y, Zhang S-L, Zheng M-Y. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction. *Nat Methods*. 2024;22:310–22.
 21. Zhang O, Lin H-T, Zhang H, Zhao H-F, Huang Y-F, Hsieh C-Y, Pan P-C, Hou T-J. Deep lead optimization: leveraging generative AI for structural modification. *J Am Chem Soc*. 2024;146:31357–70.
 22. Acharya A, Nagpure M, Roy N, Gupta V, Patranabis S, Guchhait SK. How to nurture natural products to create new therapeutics: strategic innovations and molecule-to-medicinal insights into therapeutic advancements. *Drug Discov Today*. 2024;29:1–19.
 23. Cross S, Cruciani G. FragExplorer: GRIND-based fragment growing and replacement. *J Chem Inf Model*. 2022;62:1224–35.
 24. Schneider G, Neidhart W, Giller T, Schmid G. “Scaffold-Hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed*. 1999;38:2894–6.
 25. Erlanson DA, McDowell RS, O’Brien T. Fragment-based drug discovery. *J Med Chem*. 2004;47:3463–82.
 26. Jhoti H, Williams G, Rees DC, Murray CW, Jhoti H, Williams G, Rees DC, Murray CW. The “rule of three” for fragment-based drug discovery: where are we now? *Nat Rev Drug Discov*. 2013;12:1–2.
 27. Gagare S, Patil P, Jain A. Natural product-inspired strategies towards the discovery of novel bioactive molecules. *Future J Pharm Sci*. 2024;10:1–23.
 28. Chen X, Varghese S, Zhang Z-Y, Du J-C, Ruan B-F, Baell JB, Liu X-H. Drug discovery and optimization based on the co-crystal structure of natural product with target. *Eur J Med Chem*. 2024;266:116–26.
 29. Sakano K, Furui K, Ohue M. NPGPT: natural product-like compound generation with GPT-based chemical language models. *J Supercomput*. 2024;81:1–16.
 30. Welsch ME, Snyder SA, Stockwell BR. Privileged scaffolds for library design and drug discovery. *Curr Opin Chem Biol*. 2010;14:347–61.
 31. Green H, Koes DR, Durrant JD. DeepFrag: a deep convolutional neural network for fragment-based lead optimization. *Chem Sci*. 2021;12:8036–47.
 32. Yang S, Hwang D, Lee S, Ryu S, Hwang SJ. Hit and lead discovery with explorative RL and fragment-based molecule generation. *arXiv preprint arXiv: 2110.01219* [Online]. 2023.
 33. Telepov A, Tsybin A, Khrabrov K, Yakukhnov S, Strashnov P, Zhilyaev P, Rumiantsev E, Ezhov D, Avetisyan M, Popova O, Kadurin A. FREED++: improving RL agents for fragment-based molecule generation by thorough reproduction. *arXiv preprint arXiv: 2401.09840* [Online]. 2024.
 34. Imrie F, Hadfield TE, Bradley AR, Deane CM. Deep generative design with 3D pharmacophoric constraints. *Chem Sci*. 2021;12:14577–89.
 35. Hadfield TE, Imrie F, Merritt A, Birchall K, Deane CM. Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration. *J Chem Inf Model*. 2022;62:2280–92.
 36. Eguida M, Schmitt-Valencia C, Hibert M, Villa P, Rognan D. Target-focused library design by pocket-applied computer vision and fragment deep generative linking. *J Med Chem*. 2022;65:13771–83.
 37. Zhang Z-X, Liu Q. Learning subpocket prototypes for generalizable structure-based drug design. *arXiv preprint arXiv: 2305.13997* [Online]. 2023.
 38. Shen T, Seo S, Lee G, Pandey M, Smith JR, Cherkasov A, Kim WY, Ester M. TacoGFN: target-conditioned GFlowNet for structure-based drug design. *arXiv preprint arXiv: 2310.03223* [Online]. 2024.
 39. Zhu H-M, Zhou R-Y, Cao D-S, Tang J, Li M. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nat Commun*. 2023;14:1–11.
 40. Powers AS, Yu HH, Suriana P, Koodli RV, Lu T, Paggi JM, Dror RO. Geometric deep learning for structure-based ligand design. *ACS Cent Sci*. 2023;9:2257–67.
 41. Lin H-T, Huang Y-F, Zhang O, Wu L-R, Li S-Y, Chen Z-Y, Li SZ. Functional-group-based diffusion for pocket-specific molecule generation and elaboration. *arXiv preprint arXiv: 2306.13769* [Online]. 2024.
 42. Li X-Z, Wang P-L, Fu T-F, Gao W-H, Li C-T, Shi L-L, Liu J-H. AUTODIFF: autoregressive diffusion modeling for structure-based drug design. *arXiv preprint arXiv: 2404.02003* [Online]. 2024.
 43. Yang Y-W, Ouyang S-Q, Hu X-Y, Zheng M-Y, Zhou H, Li L. Structure-based drug design via 3D molecular generative pre-training and sampling. *arXiv preprint arXiv: 2402.14315* [Online]. 2024.
 44. McNaughton AD, Bontha MS, Knutson CR, Pope JA, Kumar N. De novo design of protein target specific scaffold-based inhibitors via reinforcement learning. *arXiv preprint arXiv: 2205.10473* [Online]. 2022.
 45. Xie J-J, Chen S, Lei J-P, Yang Y-D. DiffDec: structure-aware scaffold decoration with an end-to-end diffusion model. *J Chem Inf Model*. 2024;64:2554–64.
 46. Ghorbani M, Gendele L, Beroza P, Keiser MJ. Autoregressive fragment-based diffusion for pocket-aware ligand design. *arXiv preprint arXiv: 2401.05370* [Online]. 2023.
 47. Huang L, Xu T-Y, Yu Y, Zhao P-L, Chen X-J, Han J, Xie Z, Li H-L, Zhong W-G, Wong K-C, Zhang H-T. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat Commun*. 2024;15:1–15.
 48. Zhung W, Kim H, Kim WY. 3D molecular generative framework for interaction-guided drug design. *Nat Commun*. 2024;15:1–12.
 49. Sako M, Yasuo N, Sekijima M. DiffInt: a diffusion model for structure-based drug design with explicit hydrogen bond interaction guidance. *J Chem Inf Model*. 2024;65:71–82.
 50. Xue Z, Sun C-W, Zheng W-H, Lv J-C, Liu X-G. TargetSA: adaptive simulated annealing for target-specific drug design. *Bioinformatics*. 2024;41:1–9.
 51. Chen S-C, Zhang O, Jiang C-R, Zhao H-F, Zhang X-J, Chen M-T, Liu Y, Su Q, Wu Z-X, Wang X-Y, Qu W-L, Ye Y-Y, Chai X, Wang N, Wang T-Y, An Y, Wu G-L, Yang Q-Q, Chen J-A, Xie W, Lin H-T, Li D, Hsieh C-Y, Huang Y, Kang Y, Hou T-J, Pan P-C. Deep lead optimization enveloped in protein pocket and its application in designing potent and selective ligands targeting LTK protein. *Nat Mach Intell*. 2025;7:448–58.
 52. Weller JA, Rohs R. Structure-based drug design with a deep hierarchical generative model. *J Chem Inf Model*. 2024;64:6450–63.
 53. Sommer K, Flachsenberg F, Rarey M. NAOMInext—synthetically feasible fragment growing in a structure-based design context. *Eur J Med Chem*. 2019;163:747–62.
 54. Chevillard F, Kolb P. SCUBIDOO: a large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. *J Chem Inf Model*. 2015;55:1824–35.
 55. Arús-Pous J, Patronov A, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J Cheminform*. 2020;12:1–18.
 56. Fialková V, Zhao J, Papadopoulos K, Engkvist O, Bjerrum EJ, Kogej T, Patronov A. LibINVENT: reaction-based generative scaffold decoration for in silico library design. *J Chem Inf Model*. 2021;62:2046–63.
 57. Langevin M, Minoux H, Levesque M, Bianciotto M. Scaffold-constrained molecular generation. *J Chem Inf Model*. 2020;60:5637–46.
 58. Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, Engkvist O. Reinvent 4: modern AI-driven generative molecule design. *J Cheminform*. 2024;16:1–16.
 59. Noutahi E, Gabellini C, Craig M, Lim JSC, Tossou P. Gotta be SAFE: a new framework for molecular design. *Digit Discov*. 2024;3:796–804.
 60. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. *J Chem Inf Model*. 2021;62:2064–76.
 61. Kaitoh K, Yamanishi Y. Scaffold-retained structure generator to exhaustively create molecules in an arbitrary chemical space. *J Chem Inf Model*. 2022;62:2212–25.
 62. Liao Z-R, Xie L, Mamitsuka H, Zhu S-F. Sc2Mol: a scaffold-based two-step molecule generator with variational autoencoder and transformer. *Bioinformatics*. 2023;39:1–9.
 63. Lim J, Hwang S-Y, Moon S, Kim S, Kim WY. Scaffold-based molecular design with a graph generative model. *Chem Sci*. 2020;11:153–64.

64. Liu X, Ye K, van Vlijmen HWT, IJzerman AP, van Westen GJP. DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning. *J Cheminform.* 2023;15:1–14.
65. Xu M-Y, Chen H-M. Tree-Invent: a novel multipurpose molecular generative model constrained with a topological tree. *J Chem Inf Model.* 2023;63:7067–82.
66. Maziarz K, Jackson-Flux H, Cameron P, Sirockin F, Schneider N, Stiefl N, Segler M, Brockschmidt M. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv: 2110.01219* [Online], 2024.
67. Gao Z-X, Wang X-Y, Gaines BB, Shi X-T, Bi J-B, Song M-H. Fragment-based deep molecular generation using hierarchical chemical graph representation and multi-resolution graph variational autoencoder. *Mol Inform.* 2023;42:1–15.
68. Joshi RP, Gebauer NWA, Bontha M, Khazaieli M, James RM, Brown JB, Kumar N. 3D-Scaffold: a deep learning framework to generate 3D coordinates of drug-like molecules with desired scaffolds. *J Phys Chem.* 2021;125:12166–76.
69. Xu C, Liu R-D, Huang S-H, Li W-C, Li Z, Luo H-B. 3D-SMGE: a pipeline for scaffold-based molecular generation and evaluation. *Brief Bioinform.* 2023;24:1–11.
70. Zheng S-J, Lei Z-R, Ai H-T, Chen H-M, Deng D-G, Yang Y-D. Deep scaffold hopping with multimodal transformer neural networks. *J Cheminform.* 2021;13:1–15.
71. Yu Y, Xu T-Y, Li J-W, Qiu Y-P, Rong Y, Gong Z, Cheng X-M, Dong L-M, Liu W, Li J, Dou D-F, Huang J-Z. A novel scalarized scaffold hopping algorithm with graph-based variational autoencoder for discovery of JAK1 inhibitors. *ACS Omega.* 2021;6:22945–54.
72. Hu C, Li S, Yang C-X, Chen J, Xiong Y, Fan G-S, Liu H, Hong L. ScaffoldGVAE: scaffold generation and hopping of drug molecules via a variational autoencoder based on multi-view graph neural networks. *J Cheminform.* 2023;15:1–17.
73. Torge J, Harris C, Mathis SV, Lio P. DiffHopp: a graph diffusion model for novel drug design via scaffold hopping. *arXiv preprint arXiv: 2308.07416* [Online], 2023.
74. Igashov I, Stärk H, Vignac C, Schneuing A, Satorras VG, Frossard P, Welling M, Bronstein M, Correira B. Equivariant 3D-conditional diffusion model for molecular linker design. *Nat Mach Intell.* 2024;6:417–27.
75. Imrie F, Bradley AR, Schaar MVD, Deane CM. Deep generative models for 3D linker design. *J Chem Inf Model.* 2020;60:1983–95.
76. Yang Y-Y, Zheng S-J, Su S-M, Zhao C, Xu J, Chen H-M. SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chem Sci.* 2020;11:8312–22.
77. Feng Y, Yang Y-Y, Deng W-B, Chen H-M, Ran T. SyntaLinker-Hybrid: a deep learning approach for target specific drug design. *Artif Intell Life Sci.* 2022;2:1–11.
78. Tan Y-H, Dai L-X, Huang W-F, Guo Y-F, Zheng S-J, Lei J-P, Chen H-M, Yang Y-D. DRlinker: deep reinforcement learning for optimization in fragment linking design. *J Chem Inf Model.* 2022;62:5907–17.
79. Zhang H, Huang J-C, Xie J-J, Huang W-F, Yang Y-D, Xu M-Y, Lei J-P, Chen H-M. GRELinker: a graph-based generative model for molecular linker design with reinforcement and curriculum learning. *J Chem Inf Model.* 2024;64:666–76.
80. Guo J, Knuth F, Margreitter C, Janet JP, Papadopoulos K, Engkvist O, Patronov A. Link-INVENT: generative linker design with reinforcement learning. *Digit Discov.* 2023;2:392–408.
81. Blaschke T, Arús-Pous J, Chen H-M, Margreitter C, Tyrchan C, Engkvist O, Papadopoulos K, Patronov A. REINVENT 2.0: an AI tool for de novo drug design. *J Chem Inf Model.* 2020;60:5918–22.
82. Wang Z-L, Yang L-Y, Zhao X-E. Co-crystallization and structure determination: an effective direction for anti-SARS-CoV-2 drug discovery. *Comput Struct Biotechnol J.* 2021;19:4684–701.
83. Wang Y-T, Long X-Y, Ding X, Fan S-R, Cai J-Y, Yang B-J, Zhang X-F, Luo R-H, Yang L, Ruan T, Ren J, Jing C-X, Zheng Y-T, Hao X-J, Chen D-Z. Novel nucleocapsid protein-targeting phenanthridine inhibitors of SARS-CoV-2. *Eur J Med Chem.* 2022;227:1–12.
84. Xiang Z-R, Fan S-R, Ren J, Ruan T, Chen Y, Zhang Y-W, Wang Y-T, Yu Z-Z, Wang C-F, Sun X-L, Hao X-J, Chen D-Z. Utilizing artificial intelligence for precision exploration of N protein targeting phenanthridine SARS-CoV-2 inhibitors: a novel approach. *Eur J Med Chem.* 2024;279:1–19.
85. Fortune JM, Osheroff N. Topoisomerase II as a target for anticancer drugs: when enzymes stop being nice. *Prog Nucleic Acid Res Mol Biol.* 2000;64:221–53.
86. Pogorelčnik B, Brvar M, Zajc I, Filipič M, Solmajer T, Perdih A. Monocyclic 4-amino-6-(phenylamino)-1,3,5-triazines as inhibitors of human DNA topoisomerase II α . *Bioorg Med Chem Lett.* 2014;24:5762–8.
87. Herlah B, Goričan T, Benedik NS, Grdadolnik SG, Sosič I, Perdih A. Simulation- and AI-directed optimization of 4,6-substituted 1,3,5-triazin-2(1H)-ones as inhibitors of human DNA topoisomerase II α . *Comput Struct Biotechnol J.* 2024;23:2995–3018.
88. Saini A, Patel R, Gaba S, Singh G, Gupta GD, Monga V. Adenosine receptor antagonists: recent advances and therapeutic perspective. *Eur J Med Chem.* 2022;227:1–29.
89. Qin R, Zhang H, Huang W-F, Shao Z-L, Lei J-P. Deep learning-based design and screening of benzimidazole-pyrazine derivatives as adenosine A2B receptor antagonists. *J Biomol Struct Dyn.* 2023;43:3225–41.
90. Liu M-Y, Zhang J-F, Li X-X, Wang Y-X. Research progress of DDR1 inhibitors in the treatment of multiple human diseases. *Eur J Med Chem.* 2024;268:1–21.
91. Wang Y-L, Dai Y, Wu X-W, Li F, Liu B, Li C-P, Liu Q-F, Zhou Y-Y, Wang B, Zhu M-R, Cui R-R, Tan X-Q, Xiong Z-P, Liu J, Tan M-J, Xu Y-C, Geng M-Y, Jiang H-L, Liu H, Ai J, Zheng M-Y. Discovery and development of a series of pyrazolo[3,4-d]pyridazinone compounds as the novel covalent fibroblast growth factor receptor inhibitors by the rational drug design. *J Med Chem.* 2019;62:7473–88.
92. Tan X-Q, Li C-P, Yang R-R, Zhao S, Li F, Li X-T, Chen L-F, Wan X-Z, Liu X-H, Yang T-B, Tong X-C, Xu T-Y, Cui R-R, Jiang H-L, Zhang S-L, Liu H, Zheng M-Y. Discovery of pyrazolo[3,4-d]pyridazinone derivatives as selective DDR1 inhibitors via deep learning based design, synthesis, and biological evaluation. *J Med Chem.* 2022;65:103–19.
93. Bachmaier K, Krawczyk C, Kozieradzki I, Kong Y-Y, Sasaki T, Oliveira-dos-Santos A, Mariathasan S, Bouchard D, Wakeham A, Itie A, Le J, Ohashi PS, Sarosi I, Nishina H, Lipkowitz S, Penninger JM. Negative regulation of lymphocyte activation and autoimmunity by the molecular adaptor Cbl-b. *Nature.* 2000;403:211–6.
94. Quinn TR, Giblin KA, Thomson C, Boerth JA, Bommakanti G, Braybrooke E, Chan C, Chinn AJ, Code E, Cui C, Fan Y, Grimster NP, Kohara K, Lamb ML, Ma L, Mfuh AM, Robb GR, Robbins KJ, Schimpl M, Tang H, Ware J, Wrigley GL, Xue L, Zhang Y, Zhu H, Hughes SJ. Accelerated discovery of carbamate Cbl-b inhibitors using generative AI models and structure-based drug design. *J Med Chem.* 2024;67:14210–33.
95. Lee J-H. Targeting the ATM pathway in cancer: opportunities, challenges and personalized therapeutic strategies. *Cancer Treat Rev.* 2024;129:1–14.
96. Deng D, Yang Y-X, Zou Y-R, Liu K-J, Zhang C-F, Tang M-H, Yang T, Chen Y, Yuan X, Guo Y, Zhang S-J, Si W-T, Peng B, Xu Q, He W, Xu D-G, Xiang M-L, Chen L-J. Discovery and evaluation of 3-quinoxalin urea derivatives as potent, selective, and orally available ATM inhibitors combined with chemotherapy for the treatment of cancer via goal-oriented molecule generation and virtual screening. *J Med Chem.* 2023;66:9495–518.
97. Traves PG, Murray B, Campigotto F, Galien R, Meng A, Paolo JAD. JAK selectivity and the implications for clinical inhibition of pharmacodynamic cytokine signalling by filgotinib, upadacitinib, tofacitinib and baricitinib. *Ann Rheum Dis.* 2021;80:865–75.
98. Xiang S, Song S-K, Tang H-T, Smaill JB, Wang A-Q, Xie H, Lu X-Y. TANK-binding kinase 1 (TBK1): an emerging therapeutic target for drug discovery. *Drug Discov Today.* 2021;26:2445–55.
99. Song S-K, Tang H-T, Ran T, Fang F, Tong L-J, Chen H-M, Xie H, Lu X-Y. Application of deep generative model for design of pyrrolo[2,3-d]pyrimidine derivatives as new selective TANK binding kinase 1 (TBK1) inhibitors. *Eur J Med Chem.* 2023;247:1–15.
100. Zhu Y-Y, Ouyang Z-J, Du H-J, Wang M-J, Wang J-J, Sun H-Y, Kong L-D, Xu Q, Ma H-Y, Sun Y. New opportunities and challenges of natural products research: when target identification meets single-cell multiomics. *Acta Pharm Sin B.* 2022;12:4011–39.
101. Li J-S, Gong X-Q. Harnessing pre-trained models for accurate prediction of protein-ligand binding affinity. *BMC Bioinf.* 2025;26:1–21.
102. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. *ACS Cent Sci.* 2017;3:283–93.
103. Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. Generative molecular design in low data regimes. *Nat Mach Intell.* 2020;2:171–80.

104. Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: recent advances and challenges. *Wires Comput Mol Sci*. 2022;12:1–17.
105. Wang M-Y, Li S, Wang J-K, Zhang O, Du H-Y, Jiang D-J, Wu Z-X, Deng Y-F, Kang Y, Pan P-C, Li D, Wang X-R, Yao X-J, Hou T-J, Hsieh C-Y. ClickGen: directed exploration of synthesizable chemical space via modular reactions and reinforcement learning. *Nat Commun*. 2024;15:1–18.
106. Nahal Y, Menke J, Martinelli J, Heinonen M, Kabeshov M, Janet JP, Nittinger E, Engkvist O, Kaski S. Human-in-the-loop active learning for goal-oriented molecule generation. *J Cheminform*. 2024;16:1–24.
107. Lu W, Zhang J-X, Huang W-F, Zhang Z-Q, Jia X-Y, Wang Z-Y, Shi L-L, Li C-T, Wolynes PG, Zheng S-J. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat Commun*. 2024;15:1–13.
108. Cheng K-H, Liu C, Su Q-K, Wang J, Zhang L-W, Tang Y-N, Yao Y, Zhu S-Y, Qi Y. AlphaFolding: 4D diffusion for dynamic protein structure prediction with reference and motion guidance. *arXiv preprint arXiv: 2408.12419 [Online]*, 2024
109. Wu F, Jin S-T, Jiang Y-H, Jin X-R, Tang B-W, Niu Z-M, Liu X-R, Zhang Q, Zeng X-X, Li SZ. Pre-training of equivariant graph matching networks with conformation flexibility for drug binding. *Adv Sci*. 2022;9:1–13.
110. Schuh MG, Boldini D, Bohne AI, Sieber SA. Barlow twins deep neural network for advanced 1D drug-target interaction prediction. *J Cheminform*. 2025;17:1–14.
111. Miglior L, Simone L, Podda M, Bacciu D. Towards efficient molecular property optimization with graph energy based models. *arXiv preprint arXiv: 2502.12219 [Online]*, 2025.
112. Li M-S, Zhang L, Zhu M-Z, Huang Z-L, Yu G, Fan J-Y, Chen T. Lightweight model pre-training via language guided knowledge distillation. *IEEE Trans Multimedia*. 2024;26:10720–30.
113. Perera D, Tucker JW, Brahmabhatt S, Helal CJ, Chong A, Farrell W, Richardson P, Sach NW. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*. 2018;359:429–34.
114. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. 2018;555:604–10.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.