Article published online: 2024-12-02

Recent Advancements in the Application of Artificial Intelligence in Drug Molecular Generation and Synthesis Planning

Buyong Ma[1]* Yiguo Wang[1] Xingzi Li[1] Chang Shen[1] Hao Lin[1] Chenxi Du[1] Shanlin Yang[1] Ruoqing Zeng[1] Xuyang Tang[1] Jinglei Hu[1] Yukun Yang[1] Jingwen Wang[1] Jiawei Zhu[1] Xingqian Shan[1] Yu Zhang[1] Jiaqing Hu[1]

[1] Engineering Research Center of Cell & Therapeutic Antibody (MOE), School of Pharmacy, Shanghai Jiao Tong University, Shanghai, People's Republic of China

Address for correspondence Buyong Ma, PhD, School of Pharmacy, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, People's Republic of China (e-mail: mabuyong@sjtu.edu.cn).

Pharmaceut Fronts 2024;6:e394–e405.

Abstract

The design and synthesis of drug molecules is a pivotal stage in drug development that traditionally requires significant investment in time and finances. However, the integration of artificial intelligence (AI) in drug design accelerates the identification of potential drug candidates, optimizes the drug development process, and contributes to more informed decision-making. The application of AI in molecular generation is changing the way researchers explore the chemical space and design novel compounds. It accelerates the process of drug discovery and materials science, enabling rapid exploration of the vast chemical landscapes for the identification of promising candidates for further experimental validation. The application of AI in predicting reaction products accelerates the synthesis planning process, contributes to the automation of synthetic chemistry tasks, and supports chemists in making informed decisions during drug discovery. This paper reviewed the recent advances in two interrelated areas: the application of AI in molecular generation and synthesis routes. It will provide insights into the innovative ways in which AI is transforming traditional approaches in drug development and predict its future progress in these key fields.

Keywords
- artificial intelligence
- drug screen
- molecular generation
- retrosynthesis
- deep learning

Introduction

Traditional drug discovery methods are usually associated with significant challenges including time-consuming processes, low hit rates, and a narrow focus on known targets. In comparison to the traditional drug discovery methods, the application of artificial intelligence (AI) presents a paradigm shift.[1] AI is playing a growing role in speeding up and improving drug discovery and has been widely used in pharmaceuticals and health care including, but not limited to, target prediction, virtual screening, molecular design, accelerating the identification of potential drug candidate, and optimizing their pharmacological properties, revolutionizing medical imaging and disease management,[2] enhancing operational efficiency, and minimizing downtime in health care services.

AI-driven generative models, such as deep learning-based approaches, which can be used to design new molecules with desired properties, have emerged as a new paradigm in chemical sciences.[3] This helps to create diverse chemical

received
January 16, 2024
accepted
November 2, 2024
article published online
December 2, 2024

DOI https://doi.org/10.1055/s-0044-1796647.
ISSN 2628-5088.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

libraries for drug screening. Variational autoencoders (VAEs) and generative adversarial networks (GANs) were two major models commonly used in molecular generation. VAEs learn the latent space of molecular structures and generate novel molecules by sampling from this latent space. GANs generate molecular structures by training on a dataset of known structures. They consist of a generator that creates molecules and a discriminator that evaluates how well the generated molecules resemble real ones.

AI is used to design novel drug-like molecules with desired pharmacological properties. Generative models can propose chemical structures with optimized properties such as binding affinity, solubility, and bioavailability.[4] A systematic literature review found that six AI algorithms are commonly used for *de novo* molecule generation, including evolutionary algorithms, adversarial autoencoders, VAEs, GANs, long short-term memory recurrent neural networks (RNN), and gated recurrent units.[5] Recently, the diffusion model has been widely used in molecule generations, where random noises are added into three-dimensional (3D) molecule geometries, and the desired 3D geometries are constructed by learning through a reverse process.[6]

AI models predict the outcome of chemical reactions, recommend viable reactants, and design chemical synthesis routes and conditions under which the desired product is formed,[7] which helps to generate reaction pathways of molecular synthesis. By training on databases of known reactions, these models can learn patterns and relationships between reactants and products, allowing them to predict possible outcomes of new reactions. In retrosynthetic analysis, AI suggested synthetic routes for a target molecule by considering known reactions and associated conditions to provide viable synthesis pathways for the desired compound.

Many aspects related to generative models for *de novo* drug design encompassing the categories based on molecular representations *in silico*,[8] focusing on reinforcement learning (RL),[9] incorporation of protein structure,[10] and comparing small molecule and protein generation,[11] have been documented. However, the current review highlights the transformative potential of AI across multiple facets of pharmaceuticals and health care, focusing on two interrelated approaches: AI-driven molecule generation and synthesis planning. We also systematically review molecular generations with one-dimensional (1D; SMILES strings), two-dimensional (2D; graph), and 3D ligands generations.

From Drug Screening to Artificial Intelligence-Assisted Molecular Generation

Machine learning (ML) greatly accelerates the drug discovery process in terms of virtual screening to predict pharmacokinetic properties, toxicity, bioavailability, cellular localization, and screening molecular targets and bioactivity. ML algorithms fall into two fundamental types: supervised learning and unsupervised learning. Supervised learning uses regression analysis (e.g., decision trees, random forests, support vector machines (SVMs), and artificial neural networks [AAN]) and classifier methods to train ML models specifically for datasets with active and nonactive compounds. Unsupervised algorithms classify compound datasets by identifying patterns, with examples such as Hidden Markov Models, hierarchical clustering, and k-means clustering.[12] In drug screening, ANN show great promise by efficiently filtering candidate drugs from extensive databases, simplifying the processes, and establishing relationships between multiple targets.[13] The chemical spaces in the existing databases are still limited; fortunately, AI helped molecular design by automatic generation of new drug-like molecules, offering the promise of exploring the vast chemical space.

Molecular generation methods can be divided into structure-based and ligand-based methods. The former uses high-precision structural features of the target protein pocket to provide direct guidance for optimizing the interaction between the ligand and the target, thereby driving rational compound design. The latter uses datasets of known active ligands to design effective molecules with optimal properties, which can be heavily influenced by training data.

Ligand-Based Molecular Generation and SMILES

Multiconstrained molecular generation (MCMG) utilizes knowledge distillation, combined with a conditional transformer and a QSAR (quantitative structure–activity relationship)-based RL algorithm, to satisfy multiple constraints and generate new molecules with desired pharmacological and physicochemical properties. The process involves preconditioning the generative model without destroying the output diversity. MCMG consists of three essential submodels: a prior model, a distilled model based on RNN, and an agent model. A c-transformer is trained and then distilled into RNN for subsequent application with RL. The MCMG can effectively balance the convergence speed of the molecule generation model and partially address the challenge of output diversity.[14]

De novo drug design based on the SMILES format of ligands is a convenient method because all organic compounds can be easily represented by SMILES strings. Thus, it converts ligand information processing into a sequence-processing procedure and allows for learning of grammatical rules of known compounds using various models (e.g., transformer, RNN). It is shown that in the case of BRAF inhibitor design, transformer-encoder-based generative model trained using ChEMBL's 1.6 million data sets can be fine-tuned using transfer learning and RL to design a new BRAF inhibitor with desirable activity.[15] SMILES of the ChEMBL dataset, which can also be combined with protein sequence information, can be used to generate target-binding drugs.[16]

The SMILES generators are compatible with the *de novo* generation of dual-target ligands by using two discriminators to drive molecules from the overlap of two bioactive compound distributions.[17] Even without specific 3D pocket inputs, protein–ligand interaction can incorporate the quantitative strength of common interaction types, such as van der Waals force, electrostatic interactions, and

hydrogen bonds. Integrating this energy information into a VAE framework minimizes SMILES reconstruction error and generates compounds with the desired interactions.[18] In model training, the ligand 3D grid information of atomic physicochemical properties can be combined with SMILES strings.[19]

The combination of BiLSTM (bidirectional long short-term memory) and Mol-CycleGAN (molecular cycle generative adversarial network) methods can retain molecular input information with a cycle architecture.[20] SMILES-based generative models can be generated starting from a selected core molecule and then using Monte Carlo Tree Search and a RNN to insert the generated partial SMILES into the initial core SMILES.[21]

REINVENT is a seminal molecular *de novo* design via deep RL.[22–24] It is interesting to note that REINVENT uses one-hot encoded SMILES as input (►**Fig. 1**) and uses a language-based generative model RL to maximize a reward provided by an external scoring function to optimize molecule generation.[22] The latest REINVENT 4 extends a number of functions including *de novo* design, molecule optimization, library design, R-group replacement, linker design, and scaffold hopping.[24] These functions are also included in DrugHIVE, a structure-based drug design hierarchical generative model.[25]

Molecules generated as SMILES strings are usually accompanied by invalid molecules. To address this problem, Krenn proposed self-referencing embedded strings (SELFIES), where each SELFIES corresponds to a valid molecule, even for entirely random strings.[26] TransGEM is a molecule generation model based on a transformer with gene expression data. Zhou et al used the SELFIES to construct a molecule generation model to incorporate gene expression data.[27] The study found that high attention scores obtained from the transformer model were associated with the onset of the disease, indicating the potential of these genes as disease targets.[27]

Molecular Generation with Two-Dimensional (Graphs) Molecular Representations

Graph-based molecular generation extends the description of molecular structure regarding realistic chemicals.[28,29] Fragment-based conditional molecular generation is an effective method to generate valid molecules, which can be accomplished using 1D SMILES[21] or frequently graph-based models.[30] Initial seed used for fragment-based generation can be optimized by activity-swapping methods that allow for the activation, deactivation, or retention of activity of molecular seeds.[31] To keep more global information than random fragment search, scaffold-based deep generative models are increasingly used, helping in considering stereochemical information by searching scaffold and pharmacophore constraints,[32–36] or by scaffold hopping to increase diversities.[34,36] Graph-based models can used to generate molecules with increased drug-likeness, decreased/increased size, and enhanced bioactivity.[37] Additional studies have been reported to reduce the gap between graph generative models and target-based discovery.[38]

Examples of fragment-based generators include FAME,[39] Modof,[40] and NIMO.[41] FAME treats molecules as sequences of fragments and can be combined with gene expression profiles.[39] Modof-pipe improves octanol–water partition coefficient to optimize synthetic accessibility.[40] NIMO uses two tailor-made motif extraction methods to map a molecular graph into a semantic motif sequence.[41] Drug Design based on graph-fragment molecular representation can perform multiobjective molecular optimization, including desired physicochemical properties and binding affinity scores as targets.[42]

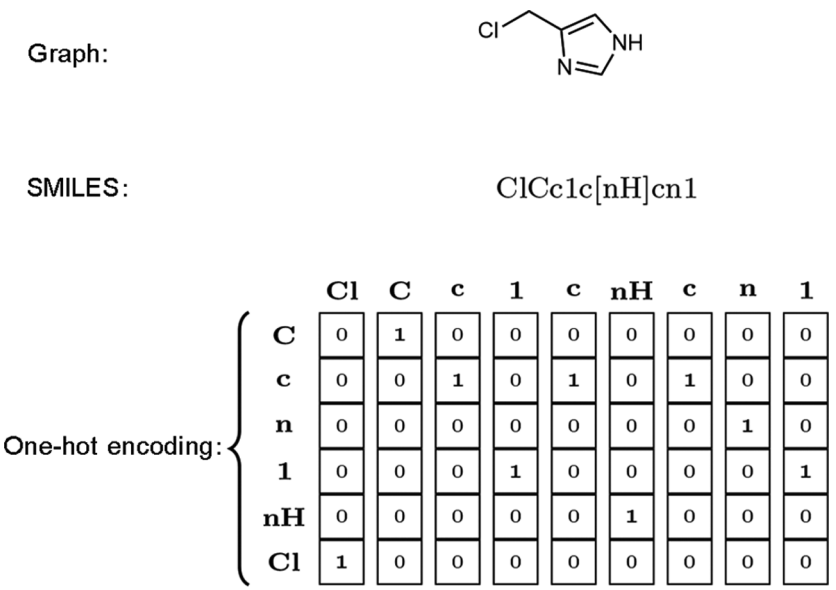

Fig. 1 Illustration of one-hot representation derived from the SMILES of 4-(chloromethyl)-1*H*-imidazole. Reproduced with permission from Olivecrona et al.[22]

Many fragment-based molecular generation models use Monte Carlo tree search (MCTS) to find an optimal attached point for extended fragment growth.[43–46] MCTS combines the standards of Monte Carlo strategies with tree-primarily based search techniques that sample and explore only promising areas of the targeted area. VGAE-MCTS is a molecular generative model that combines the variational graph autoencoder and MCTS.[43] Mothra and AlphaDrug also use MCTS as a conditional molecular generation algorithm.[44,45]

Existing graph-based deep generative models can be easily extended to 3D representations of molecules and target pharmacophores.[47] 3D-based models are more efficient in exploring chemical space in comparison to 2D methods.[48] The graph information embedded in relative coordinates also helps to encode the 3D structure of a molecular, thus satisfying the requirements of translation and rotation invariance.[49]

Structure-Based Molecular Generation and 3D Information of Protein Binding Sites

Great advancements have been made in *de novo* drug design using 3D deep generative models.[50] Generated molecular properties and protein binding affinity often depend on the environment in which the protein binds to the ligand. Protein binding pocket can be represented by atomic density grids,[51] electrostatic environments,[52] or experimental electron density directly.[53] In the RELATION model, geometric features of the desired protein–ligand complexes were extracted and transferred to a latent space for generation.[54]

An important aim of incorporating protein structure into molecule generation is to maximize the predicted on-target binding affinity of generated molecules.[10,55] Specific physics-based features including the binding mechanism between a receptor and a ligand,[56] or drug–target interaction, can be described in model training.[52,53] For example, four kinds of atomic interactions including π–π interaction, cation–π interaction, hydrogen bond interaction, and halogen bond interaction, were tested using learnable vector embeddings with a diffusion model.[57] The ligand–protein interaction can also be converted to fingerprint as constraints.[58]

Indeed, diffusion-based generative models have proved to be a powerful tool.[59] PILOT is a diffusion-based *de novo* ligand generation that combines pocket conditioning with large-scale pretraining and property guidance. For a given pocket of proteins, the generated molecules have higher binding affinity while maintaining high synthetic accessibility.[60] PMDM is a dual diffusion model consisting of a conditional equivariant diffusion model with both local and global molecular dynamics.[61]

ResGen is a SE(3)-equivariant conditional generative model that generates 3D molecules based on the structure of the protein pocket.[62] The model employs a parallel multiscale modeling strategy and a two-level autoregression protocol, which is capable of capturing higher-level interaction between protein targets and ligands with better computational efficiency. The molecules generated by ResGen can bind tightly to previously unseen protein pockets of therapeutic relevance, have potentially enhanced drug-likeness and ease of synthesized properties, and are closely similar to the known active compounds. Notably, ResGen could be used for conformation generation and analysis.[62]

Since the emergence of 3D molecular generation models, most methods have conditioned on the target structure, thereby neglecting interaction information related to complex molecule conceptualization and stability. In SurfGen, inspired by the simple lock–key mode, protein surface channels are used as protein representation.[63] To stimulate complementarity between small molecules and protein pockets, topology learning was subsequently performed via a Geodesic-Graph Neural Network (Geodesic-GNN). SurfGen has the highest performance in docking and scoring compared with other methods, e.g., GraphBP and Pocket2Mol, and can generate molecules with highly similar electron distribution and shape to the original ligand. SurfGen's high sensitivity to pocket structures provides an effective solution for drug resistance.

Deep learning-based molecular generation methods produce some biases related to the ligands in the training sets, which restrict their application to data with limited biological activity. To achieve structure-based 3D molecular design, a new network architecture—Ligand neural network (L-Net), is used for end-to-end 3D molecular construction.[46] L-Net is based on a graph convolutional neural network and is trained using molecular structures extracted from the ChEMBL database, which allows for the generation of drug-like molecules with high-quality 3D conformations. Combining L-Net and MCTS (Monte Carlo tree search) algorithm, DeepLigBuilder is developed to achieve *de novo* drug design based on target structures, which allows direct manipulation of 3D molecular structure while optimizing the topology and 3D structure of the molecules in the binding pocket.

Molecular design still faces many challenges.[64] The performances of these models may be unsatisfactory when generating a large number of molecules with a lack of diversity.[8] How to synthesize strange molecules generated by AI, is an open question, and much work has been devoted to obtaining synthesizable molecules, e.g., by selecting reactants from commercially available compounds and constructing a synthesis route as a tree of reaction template.[65] ChemistGA combines deep learning with a genetic algorithm to enhance the accessibility and success of synthesizing molecules with desired properties.[66]

Artificial Intelligence Optimization of the Reaction Product and Reaction Conditions

Artificial Intelligence Prediction of Reaction Outcome

Accurately predicting the outcome of organic reactions is the core of organic synthesis in chemical drug development. Usually, this depends on the chemists' experience and past reaction data, which is largely driven by intuition. In contrast to traditional methods, AI algorithms can be trained with a large number of reaction precedents in literature covering a wide range of reaction types. AI can provide possible reaction products with a high degree of accuracy, and much faster

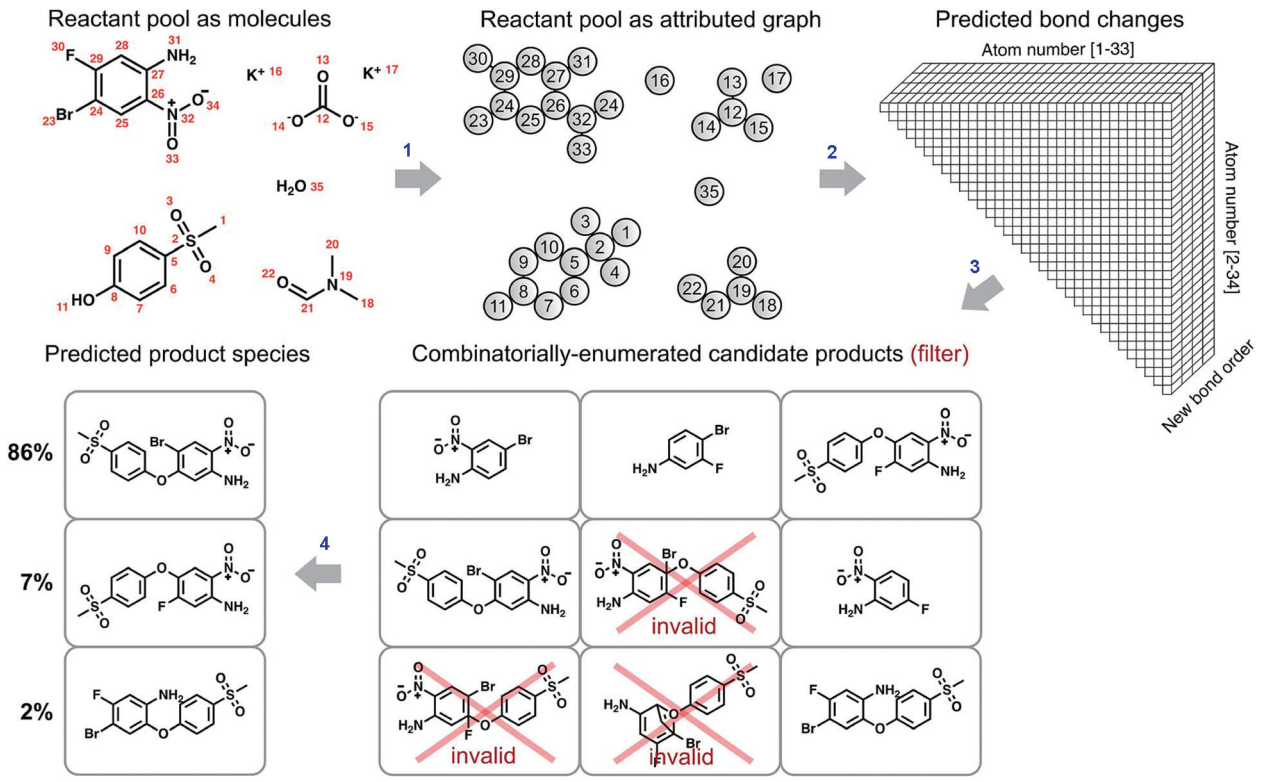

Fig. 2 Weisfeiler–Lehman network model for predicting probability distribution in reaction product mixtures. Reproduced with permission from Coley et al.[69]

than humans after considering various reaction conditions and physicochemical parameters. The algorithm can be template-free, e.g., predictions are automatically inferred from the characteristics of reactant, reagent, and product present in the data set, i.e., looking for correlations between the presence and absence of common chemical motifs.[67] In another aspect, Chen and Jung proposed a generalized-template-based GNN for accurate prediction of organic reactivity.[68] The network is based on a generalized reaction template that catches organic reactivity from the net changes in electron configuration between reactants and products.

The chemical structure figure provides a natural way to describe the structure of molecules; nodes correspond to atoms and edges to bonds. Convolutional neural networks use graph theory methods to understand chemical reactivity and predict reaction results through graph editing. In 2019, Coley and colleagues used graph-convolutional neural networks to predict the probability distribution of a mixture of reaction products.[69] As illustrated in ►**Fig. 2**, the reactant molecules, including building blocks, catalysts, bases, ligands, and solvents, are represented as atomic maps before encoding the data. Deep learning methods evaluate the probability of chemical bond recombination, predict the most likely changes, and generate a set encompassing all potential products through enumeration. Subsequently, a new convolutional neural network reallocates the initially predicted products according to the rules of the chemical valence state to establish a probability distribution. According to statistical models, the molecule with the highest probability corresponds to the primary product. The method incorporates solvent information and descriptions of all relevant species as molecular maps for atomic mapping, which significantly improves performance and enhances model interpretability. In more than 85% of cases, the main reaction products can be accurately identified, with each molecule computed in just 100 ms.

In ML, a random forest serves as a classifier consisting of multiple decision trees whose output categories are determined by the mode of individual tree outputs. Its versatility extends to handling classification, regression, and dimensionality reduction problems. Notably, random forests exhibit robustness against outliers and noise, showcasing superior predictive and classification performance compared with independent decision trees. The main advantages of random forests include: (1) generating highly accurate classifiers for various data types; (2) handling a substantial number of input variables; and (3) assessing the importance of variables in category determination. Despite these advantages, the application of random forest algorithms in chemical synthesis still faces historical challenges. The complexity of implementation, particularly for nonprofessionals, posed an obstacle. In addition, the "curse of dimensions," where data requirement grows exponentially with the number of dimensions studied, added to the complexity. This challenge is particularly pronounced in the multidimensional nature of chemical structure and reactivity, making it difficult to collect sufficient, complete, and consistent data from databases to implement algorithms.[70] Fortunately, the database established through high-throughput experiments has made it possible to predict reaction products through random

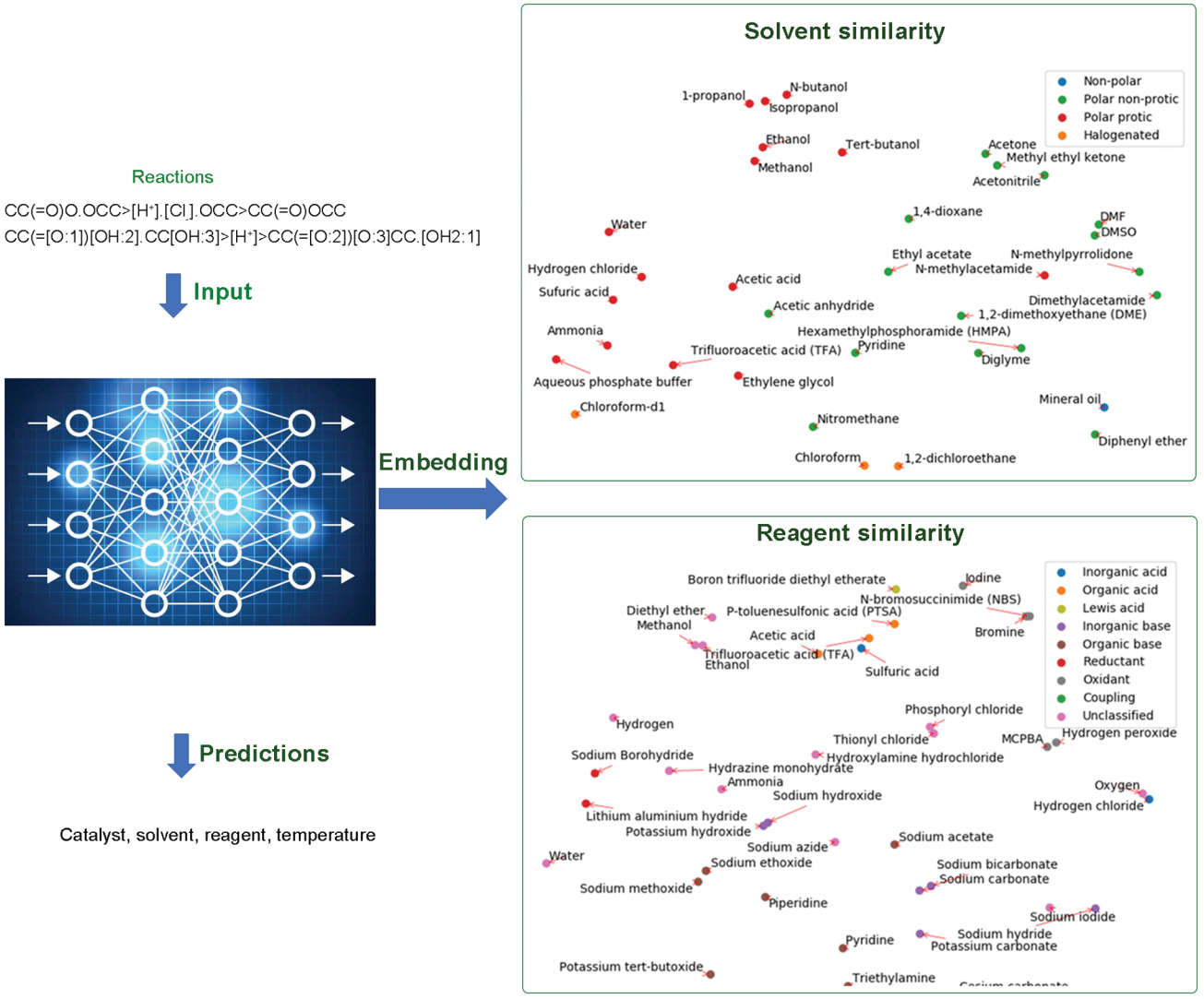

Fig. 3 Machine learning models can predict the conditions of organic synthesis reactions and quantify the similarity of solvents and reagents. (Reproduced with permission from Gao et al.[76] This is an unofficial adaptation of an article that appeared in an ACS publication. ACS has not endorsed the content of this adaptation or the context of its use.)

forests.[71] In 2018, Ahneman and colleagues trained a random forest algorithm using high-throughput datasets to predict which specific palladium catalyst is most tolerant to imidazole during C–N bond formation.[72] These predictions also help guide the analysis of catalyst inhibition mechanisms.

Ross et al proposed MoLFormer, an efficient transformer encoder model for predicting a variety of different molecular properties, which was trained on SMILES sequences of 1.1 billion unlabeled molecules from the PubChem and ZINC datasets.[73] Yoshikawa et al performed CLAIRIFY for automation of experiments in a chemistry lab using general-purpose robot manipulators and natural language commands. The large language models (LLMs) make chemical reactions more scientific, reasonable, effective, and practical, and provide stronger support and guarantee for the development and application of chemistry.[74]

In summary, AI has significant advantages over traditional methods in predicting reaction products, as it can predict the main products of a chemical reaction in a very short time and with a high degree of accuracy. The disadvantage, however, is that when it comes to predicting a reaction, a specific model needs to be established for that reaction, and the higher the accuracy, the larger the data required to train the model.

Application of Artificial Intelligence in Predicting Reaction Conditions

While AI excels at predicting reaction outcomes, it remains a challenge to experimentally verify computer-generated predictions, especially to determine the reaction conditions. The reaction conditions include the chemical environment (catalyst, reagent, solvent) and operational parameters (temperature, pressure, etc.). Different reaction conditions often produce different results. Thus, employing AI to predict and optimize these conditions can help improve the precision and success rate of reaction predictions.[75]

Gao et al discussed how ML can be used to predict fitness bars for organic reactions (►**Fig. 3**).[76] The authors pointed out major limitations of existing methods, including the inability to accurately predict complete reaction conditions, the lack of consideration of chemical background and

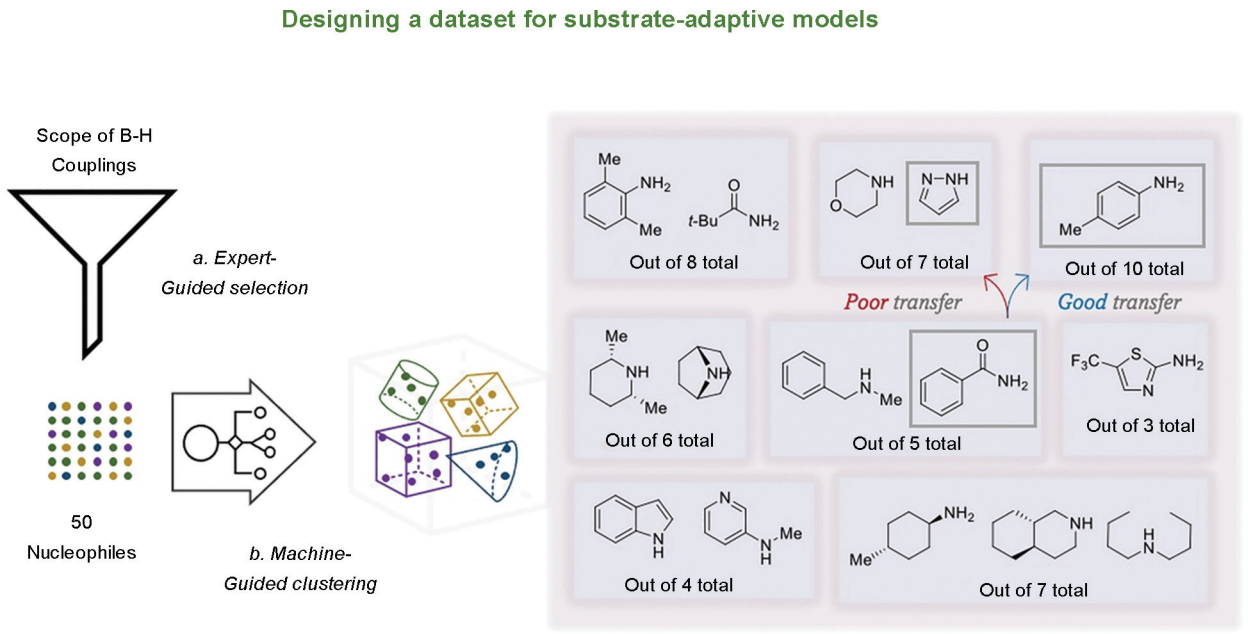

Fig. 4 Representative scope of nitrogen nucleophiles for the B–H coupling reaction and comparison to other validated ML studies on B-H couplings. Reproduced with permission from Rinehart et al.[78]

temperature compatibility, and the lack of large-scale reaction data machine-readable data. To overcome these challenges, the authors developed a neural network-based model that trained approximately 10 million reactions on Reaxys to predict appropriate reaction conditions for organic conversion. The advantages of the model include coverage of a wide range of organic reactions, the ability to predict factors of reaction conditions, and the ability to quantify the similarity of reaction conditions. However, the model has some limitations, such as a limited number of predictions and a limited ability to predict unusual situations. The potential applications of the model are also mentioned in the text, including route screening and prioritization at the path level.

Amar et al developed a hybrid mechanical-machine learning method for solvent selection in process development.[77] They used a library of 459 solvents and calculated 12 conventional molecular descriptors, two reaction-specific descriptors, and additional descriptors based on the screening charge density. The method combines physically meaningful solvent descriptors with a Gauss process-based algorithm to find solvents that are more favorable for asymmetric hydrogenation, and better than intuitively selected solvents in terms of conversion and enantiomer. In addition, automated ML workflow is successful for solvent selection. However, this approach requires a large amount of data support and needs to be complemented by proxy models with statistical predictive capabilities. Continuing to develop bridges between chemical information and data-intensive ML methods makes a lot of sense and promises to save time and resources for process chemists.

Rinehart et al developed an ML tool to predict substrate-adaptative conditions for palladium-catalyzed C–N coupling reactions.[78] The neural network model actively learns a wide range of C–N coupling reactions by designing an experimental data set. A challenge model using a neural network model was used in experimental validation and successfully isolated 10 products from a series of samples in over 85% yields. In addition, the prediction ability of the model is gradually improving with the continuous accumulation of data (►**Fig. 4**).

Gong et al introduced DeepReac+, a computational framework designed for predicting chemical reactions and determining optimal reaction conditions (►**Fig. 5**).[79] DeepReac+ includes the DeepReac model and sampling strategy and offers a robust solution. The DeepReac Model is a graph-neural network-based model that specializes in chemical reaction representation learning. It takes 2D molecular structures as inputs, adapting seamlessly to a variety of prediction tasks, including yield and stereoselectivity. Graph Attention Network, serving as its core, facilitates modeling interactions among reaction components. The sampling strategy is a key element of the DeepReac+ framework, employing two strategies: diversity sampling and adversary sampling. These strategies select informative experimental data during model training to improve model performance and cost efficiency. DeepReac+ efficiently predicts chemical reaction outcomes and identifies optimal reaction conditions by combining the DeepReac model with an active learning strategy. This integration positions DeepReac+ as a valuable AI tool in chemical synthesis.

Shields et al presented a Bayesian reaction optimization framework along with an open-source software tool.[80] This tool empowers chemists to seamlessly incorporate state-of-the-art optimization algorithms into their everyday experiments, facilitating an enhanced and user-friendly approach to reaction optimization. A large baseline data set of palladium-catalyzed direct arylation reactions is collected, and Bayes optimization and human decision-making in reaction optimization are systematically investigated. Bayesian optimization is applied to two real optimization works (Mitsunobu reaction

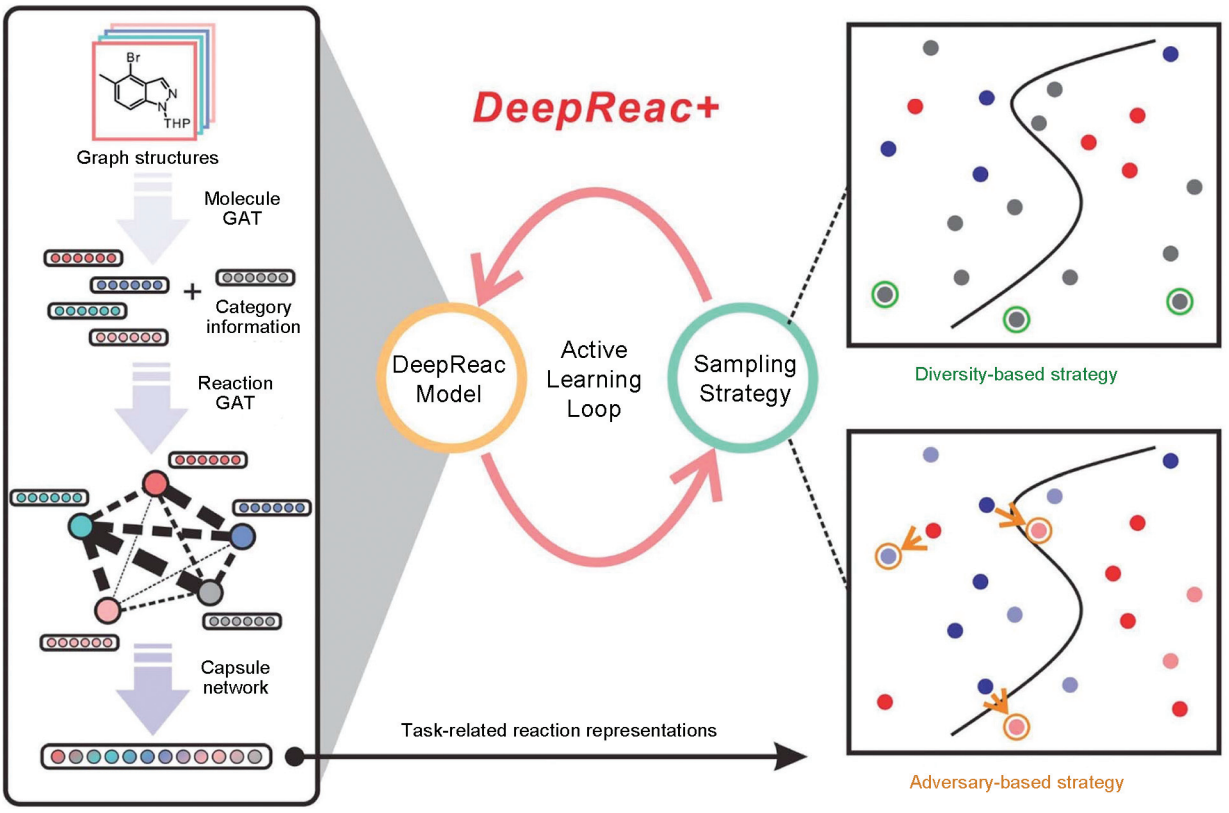

Fig. 5 Schematic workflow of the DeepReac+ framework. Reproduced with permission from Gong et al.[79]

and defluorination reaction). Its main advantages are higher average optimization efficiency and better consistency, highlighting the potential of Bayesian optimization and allowing us to make better-informed, data-driven decisions about which experiments to run, ultimately leading to more efficient synthesis of functional chemicals.

Burger et al reported a method for autonomous experimental search using mobile robots.[81] A mobile robot took center stage in the quest for enhanced photocatalysts for water-based hydrogen production. Over 8 days, the mobile robot executed a remarkable 688 experiments, using a batch Bayesian search theory algorithm within a 10-variable experimental space. This innovative approach greatly accelerated the exploration of improved photocatalysts, showcasing the efficiency and potential of robotic systems in accelerating experimentation processes. The results showed that this autonomous search identified the photocatalyst mixture that was six times more active than the original formulation, selected the beneficial component, and eliminated the negative component. Autonomous experiment search using mobile robots is flexible, efficient, and safe, but the Bayesian optimization algorithm is still somewhat blind and requires a large amount of initial investment. Therefore, it is still a good hope that robots will replace humans in experimental operations.

Currently, the prediction of reaction conditions (RCs) using a DL framework is hindered by several factors, including (1) the lack of a standardized dataset for benchmarking, (2) the lack of a general prediction model with powerful representation, and (3) the lack of interpretability. To address these issues, we first created two standardized RC datasets covering a broad range of reaction classes and then proposed a powerful and interpretable Transformer-based RC predictor named Parrot.

There are several factors affecting the prediction of reaction condition,[82] including the lack of a general prediction model with powerful representation and the lack of a standardized dataset for benchmarking. The lack of interpretability is common for most ML models. Based on a self-attention mechanism, the Transformer may boost prediction accuracy and provide interpretability, as demonstrated by interpretable Transformer-based reaction condition predictor Parrot[82] and Molecular Transformer.[83] Meanwhile, Relational Graph Convolutional Networks may also provide accurate multilabel classification solutions for prediction of reaction conditions.[84]

Application of Artificial Intelligence in Reaction Yield Prediction

With the growing abundance of molecular property datasets and reaction datasets, coupled with advancements in computing power, the application of ML technology in reactivity prediction has garnered significant attention.[85,86] Notably, Reymond and colleagues[86] showcased the extension of Natural Language Processing architectures, particularly the Transformer-based bidirectional encoder representations from transformers (BERT), for predicting reaction yields based on SMILES representations of reactants. Their work

involved fine-tuning a BERT encoder with a regression layer, pretrained using a masked language to model the loss of chemical reactions, resulting in high-quality yield predictions. The model was trained on two distinct datasets: one from high-throughput experimentation (HTE) and another from patent datasets. This trained model demonstrated its capabilities to predict a variety of reactions, including Buchwald–Hartwig and Suzuki–Miyaura, including data from the U.S. Patent and Trademark Office (USPTO) dataset. It is worth noting that the HTE and USPTO datasets differ significantly in content and quality, with the former covering a specific chemical reaction region and providing high-quality data, while the latter spans a broader reaction space with noisy and sparse data.[86] Additional studies have highlighted the limitations of this dataset's suitability for reaction yield prediction. Saebi et al found that using data from electronic laboratory notebooks (ELNs) to train attributed GNNs does not lead to a predictive model, contrary to the initial expectation that the ELNs could provide less biased, large datasets.[87]

Yield prediction methods vary using techniques such as one-hot encoding of reactants, tandem molecular fingerprints, or computational chemical descriptors. Probst et al proposed a differential reaction fingerprint (DRFP) for reaction searching and categorization as well as yield prediction.[88] The DRFP algorithm takes a reaction SMILES as an input and creates a binary fingerprint based on the symmetric difference of two sets containing circular n-grams.

Glorius'group introduced a structure-based ML platform with diverse applications in organic chemistry,[89] to achieve generality in molecular representation, they developed an input based on a multifingerprint feature. This approach applies to a variety of problem sets. Initially, it was able to accurately predict the molecular properties of diverse molecular arrays. Then, the platform successfully predicted reaction outcomes, including stereoselectivity and yield, for previously evaluated experimental datasets using problem-specific descriptor models. In a final application, the platform showed effective correlations when applied to the systematic analysis of a high-throughput dataset, showcasing its practical utility in structure-based modalities.

Reymond's group[90] used a natural language processing architecture to predict response properties based on a text-based response representation. Using an encoder-transformer model paired with a regression layer, they achieved excellent predictive performance on two high-throughput experimental reaction sets. However, when analyzing yields from the USPTO dataset, they observed differences in distribution based on mass scale. To obtain a high-quality generic reaction yield dataset, Yin et al curated a generic reaction yield dataset containing information on 12 reaction categories and reaction conditions.[91] Subsequently, using BERT-based reaction yield predictor, they found that contrastive learning based on reaction conditions enhances the sensitivity of the model to reaction conditions.

ML models using quantum chemical calculations were trained to predict the transition state and yield in copper-catalyzed P–H insertion reactions.[92] The transition state was identified by analyzing 120 experimental data points using density functional theory. Subsequently, an ML algorithm was applied to analyze the 16 descriptors derived from the transition states to predict product yields. Among the algorithms investigated, SVM had the highest prediction accuracy of 97%, with a correlation of over 80% in leave-one-out cross-validation. Sensitivity analysis was performed for each descriptor and the reaction mechanism was thoroughly examined to enhance the understanding of transition state characteristics. Matsubara's group used ML methods to build a multiple linear regression model based on batch reaction data of 29 substrates to predict the Wittig methylene reaction rate diagram of any aldehyde and diiodomethane.[93] The predicted profile allows the simultaneous determination of the highest achievable yield and the shortest reaction time. This can be interpreted as the residence time required to reach the maximum yield of the methylation of diiodomethane in a flow microreactor.

Conclusion

AI offers new opportunities for the design of innovative chemical drugs, and it has changed the traditional research paradigm of medicinal chemistry by designing and generating small molecules in a more efficient, smarter, and more precise way, and increasing their potential to become drugs.

For a given chemical reaction, AI has been developed to predict reaction products, and reaction yields, and to optimize reaction conditions. However, accurate prediction of reaction products is often dependent on the amount of data required to train the model. For the optimization of reaction conditions, there are difficulties such as the inability to accurately predict complete reaction conditions, the lack of consideration of chemical background and temperature compatibility, and the lack of machine-readable data for large-scale reaction data. The neural network-based model developed by the researchers, as well as the use of mobile robots for autonomous experimental searches, opens up additional possibilities for more comprehensively determining optimal chemical reaction conditions. Feature learning methods such as language model (LM) and GNN show good promise in chemical reaction yield prediction. On this basis, researchers have proposed structure-based ML platforms or means incorporating quantum chemical computing to accurately predict the highest yields and shortest reaction times of chemical reactions.

In small-molecule drug design, AI techniques are utilized to generate molecules with potential biological activity. Structure- and ligand-based molecular generation models offer the possibility of efficient molecular discovery. However, many challenges remain regarding the structural diversity of the generated molecules and the ability of current molecular generation models to generate large molecules. In terms of drug screening, AI realizes the

validation of drug targets and the optimal design of drug structures faster than conventional drug screening techniques based on traditional multidisciplinary. Different algorithms, as well as predictive models, have been used to evaluate the physicochemical properties as well as in vivo activity and toxicity of small molecule drugs. These technological innovations can significantly reduce the time required for new drug discovery.

Funding
This work was supported by the Natural Science Foundation of China (Grant No. 32171246) and the Shanghai Municipal Government Science Innovation (Grant No. 21JC1403700).

Conflict of Interest
None declared.

References

1 Pantelidis P, Spartalis M, Zakynthinos G, et al. Artificial intelligence: the new "fuel" to accelerate pharmaceutical development. Curr Pharm Des 2022;28(26):2127–2128
2 Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK. Recurrent residual U-Net for medical image segmentation. J Med Imaging (Bellingham) 2019;6(01):014006
3 Anstine DM, Isayev O. Generative models as an emerging paradigm in the chemical sciences. J Am Chem Soc 2023;145(16): 8736–8750
4 Sousa T, Correia J, Pereira V, Rocha M. Generative deep learning for targeted compound design. J Chem Inf Model 2021;61(11): 5343–5361
5 Martinelli DD. Generative machine learning for *de novo* drug discovery: a systematic review. Comput Biol Med 2022; 145:105403
6 Choi S, Seo S, Kim BJ, Park C, Park S. PIDiff: Physics informed diffusion model for protein pocket-specific 3D molecular generation. Comput Biol Med 2024;180:108865
7 Klucznik T, Mikulak-Klucznik B, McCormack MP, et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. Chem 2018;4:522–532
8 Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models. Brief Bioinform 2021;22(06):bbab344
9 Tong X, Liu X, Tan X, et al. Generative models for *de novo* drug design. J Med Chem 2021;64(19):14011–14027
10 Thomas M, Bender A, de Graaf C. Integrating structure-based approaches in generative molecular design. Curr Opin Struct Biol 2023;79:102559
11 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. Brief Bioinform 2024;25(04):bbae338
12 Rácz A, Bajusz D, Héberger K. Consistency of QSAR models: correct split of training and test sets, ranking of models and performance parameters. SAR QSAR Environ Res 2015;26(7-9):683–700
13 Parvatikar PP, Patil S, Khaparkhuntikar K, et al. Artificial intelligence: machine learning approach for screening large database and drug discovery. Antiviral Res 2023;220:105740
14 Wang J, Hsieh K, Wang M, et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. Nat Mach Intell 2021; 3:914–922
15 Yang L, Yang G, Bing Z, et al. Transformer-based generative model accelerating the development of novel braf inhibitors. ACS Omega 2021;6(49):33864–33873
16 Chen Y, Wang Z, Wang L, et al. Deep generative model for drug design from protein target sequence. J Cheminform 2023;15(01): 38
17 Lu F, Li M, Min X, Li C, Zeng X. *De novo* generation of dual-target ligands using adversarial training and reinforcement learning. Brief Bioinform 2021;22(06):bbab333
18 Ozawa M, Nakamura S, Yasuo N, Sekijima M. IEV2Mol: molecular generative model considering protein-ligand interaction energy vectors. J Chem Inf Model 2024;64(18):6969–6978
19 Song T, Ren Y, Wang S, et al. DNMG: deep molecular generative model by fusion of 3D information for *de novo* drug design. Methods 2023;211:10–22
20 Zhang C, Xie L, Lu X, Mao R, Xu L, Xu X. Developing an improved cycle architecture for AI-based generation of new structures aimed at drug discovery. Molecules 2024;29(07):1499
21 Erikawa D, Yasuo N, Sekijima M. MERMAID: an open source automated hit-to-lead method based on deep reinforcement learning. J Cheminform 2021;13(01):94
22 Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular *de-novo* design through deep reinforcement learning. J Cheminform 2017; 9(01):48
23 Blaschke T, Arús-Pous J, Chen H, et al. REINVENT 2.0: an AI tool for *de novo* drug design. J Chem Inf Model 2020;60(12):5918–5922
24 Loeffler HH, He J, Tibo A, et al. Reinvent 4: modern AI-driven generative molecule design. J Cheminform 2024;16(01):20
25 Weller JA, Rohs R. Structure-based drug design with a deep hierarchical generative Model. J Chem Inf Model 2024;64(16): 6450–6463
26 Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Mach Learn Sci Technol 2020;1:045024
27 Zhou Y, Wang Z, Huang Z, et al. *In silico* prediction of ocular toxicity of compounds using explainable machine learning and deep learning approaches. J Appl Toxicol 2024;44(06):892–907
28 Mercado R, Bjerrum EJ, Engkvist O. Exploring graph traversal algorithms in graph-based molecular generation. J Chem Inf Model 2022;62(09):2093–2100
29 Lee M, Min K. MGCVAE: multi-objective inverse design via molecular graph conditional variational autoencoder. J Chem Inf Model 2022;62(12):2943–2950
30 Gao Z, Wang X, Blumenfeld Gaines B, Shi X, Bi J, Song M. Fragment-based deep molecular generation using hierarchical chemical graph representation and multi-resolution graph variational autoencoder. Mol Inform 2023;42(05):e2200215
31 Kang SG, Morrone JA, Weber JK, Cornell WD. Analysis of training and seed bias in small molecules generated with a conditional graph-based variational autoencoder horizontal line insights for practical AI-driven molecule generation. J Chem Inf Model 2022; 62(04):801–816
32 Xu T, Wang M, Liu X, et al. A scaffold-based deep generative model considering molecular stereochemical information. Mol Inform 2022;41(12):e2200088
33 Langevin M, Minoux H, Levesque M, Bianciotto M. Scaffold-constrained molecular generation. J Chem Inf Model 2020;60(12): 5637–5646
34 Zheng S, Lei Z, Ai H, Chen H, Deng D, Yang Y. Deep scaffold hopping with multimodal transformer neural networks. J Cheminform 2021;13(01):87
35 Xu C, Liu R, Huang S, Li W, Li Z, Luo HB. 3D-SMGE: a pipeline for scaffold-based molecular generation and evaluation. Brief Bioinform 2023;24(06):bbad327
36 Hu C, Li S, Yang C, et al. ScaffoldGVAE: scaffold generation and hopping of drug molecules via a variational autoencoder based on multi-view graph neural networks. J Cheminform 2023;15(01): 91
37 Atance SR, Diez JV, Engkvist O, Olsson S, Mercado R. *De novo* drug design using reinforcement learning with graph-based deep generative models. J Chem Inf Model 2022;62(20):4863–4872

38 Hu F, Wang D, Huang H, Hu Y, Yin P. Bridging the gap between target-based and cell-based drug discovery with a graph generative multitask model. J Chem Inf Model 2022;62(23):6046–6056
39 Pham TH, Xie L, Zhang P. FAME: fragment-based conditional molecular generation for phenotypic drug discovery. Proc SIAM Int Conf Data Min 2022;2022:720–728
40 Chen Z, Min MR, Parthasarathy S, Ning X. A deep generative model for molecule optimization via one fragment modification. Nat Mach Intell 2021;3(12):1040–1049
41 Shen X, Zeng T, Chen N, Li J, Wu R. NIMO: a natural product-inspired molecular generative model based on conditional transformer. Molecules 2024;29(08):1867
42 Mukaidaisi M, Vu A, Grantham K, Tchagang A, Li Y. Multi-objective drug design based on graph-fragment molecular representation and deep evolutionary learning. Front Pharmacol 2022; 13:920747
43 Iwata H, Nakai T, Koyama T, Matsumoto S, Kojima R, Okuno Y. VGAE-MCTS: a new molecular generative model combining the variational graph auto-encoder and Monte Carlo tree search. J Chem Inf Model 2023;63(23):7392–7400
44 Suzuki T, Ma D, Yasuo N, Sekijima M. Mothra: multiobjective *de novo* molecular generation using monte carlo tree search. J Chem Inf Model 2024;64(19):7291–7302
45 Qian H, Lin C, Zhao D, Tu S, Xu L. AlphaDrug: protein target specific *de novo* molecular generation. PNAS Nexus 2022;1(04):pgac227
46 Li Y, Pei J, Lai L. Structure-based *de novo* drug design using 3D deep generative models. Chem Sci (Camb) 2021;12(41): 13664–13675
47 Imrie F, Hadfield TE, Bradley AR, Deane CM. Deep generative design with 3D pharmacophoric constraints. Chem Sci (Camb) 2021;12(43):14577–14589
48 Papadopoulos K, Giblin KA, Janet JP, Patronov A, Engkvist O. *De novo* design with deep generative models based on 3D similarity scoring. Bioorg Med Chem 2021;44:116308
49 Xu M, Huang W, Xu M, Lei J, Chen H. 3D conformational generative models for biological structures using graph information-embedded relative coordinates. Molecules 2022;28(01):321
50 Xie W, Wang F, Li Y, Lai L, Pei J. Advances and challenges in *de novo* drug design using three-dimensional deep generative models. J Chem Inf Model 2022;62(10):2269–2279
51 Ragoza M, Masuda T, Koes DR. Generating 3D molecules conditional on receptor binding sites with deep generative models. Chem Sci (Camb) 2022;13(09):2701–2713
52 Xu M, Ran T, Chen H. *De novo* molecule design through the molecular generative model conditioned by 3D information of protein binding sites. J Chem Inf Model 2021;61(07):3240–3254
53 Zhung W, Kim H, Kim WY. 3D molecular generative framework for interaction-guided drug design. Nat Commun 2024;15(01):2688
54 Wang M, Hsieh CY, Wang J, et al. RELATION: a deep generative model for structure-based *de novo* drug design. J Med Chem 2022; 65(13):9478–9492
55 Li S, Hu C, Ke S, et al. LS-MolGen: ligand-and-structure dual-driven deep reinforcement learning for target-specific molecular generation improves binding affinity and novelty. J Chem Inf Model 2023;63(13):4207–4215
56 Sagar D, Risheh A, Sheikh N, Forouzesh N. Physics-guided deep generative model for new ligand discovery. ACM BCB 2023;2023: 10.1145/3584371.3613067
57 Wu P, Du H, Yan Y, Lee TY, Bai C, Wu S. Guided diffusion for molecular generation with interaction prompt. Brief Bioinform 2024;25(03):bbae174
58 Zhang J, Chen H. *De novo* molecule design using molecular generative models constrained by ligand-protein interactions. J Chem Inf Model 2022;62(14):3291–3306
59 Nakata S, Mori Y, Tanaka S. End-to-end protein-ligand complex structure generation with diffusion-based generative models. BMC Bioinformatics 2023;24(01):233
60 Cremer J, Le T, Noé F, Clevert DA, Schütt KT. PILOT: equivariant diffusion for pocket-conditioned *de novo* ligand generation with multi-objective guidance via importance sampling. Chem Sci (Camb) 2024;15(36):14954–14967
61 Huang L, Xu T, Yu Y, et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. Nat Commun 2024;15(01):2657
62 Zhang O, Zhang J, Jin J, et al. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modeling. Nat Mach Intell 2023;5:1020–1030
63 Zhang O, Wang T, Weng G, et al. Learning on topological surface and geometric structure for 3D molecular generation. Nat Comput Sci 2023;3(10):849–859
64 Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci 2022;12(05): e1608
65 Nguyen DH, Tsuda K. Generating reaction trees with cascaded variational autoencoders. J Chem Phys 2022;156(04):044117
66 Wang J, Wang X, Sun H, et al. ChemistGA: a chemical synthesizable accessible molecular generation algorithm for real-world drug discovery. J Med Chem 2022;65(18): 12482–12496
67 Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5(09):1572–1583
68 Chen S, Jung Y. A generalized-template-based graph neural network for accurate organic reactivity prediction. Nat Mach Intell 2022;4:772–780
69 Coley CW, Jin W, Rogers L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci (Camb) 2018;10(02):370–377
70 Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021;4(01):153
71 Collins KD, Gensch T, Glorius F. Contemporary screening approaches to reaction discovery and development. Nat Chem 2014;6(10):859–871
72 Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C-N cross-coupling using machine learning. Science 2018;360(6385):186–190
73 Ross J, Belgodere B, Chenthamarakshan V, et al. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 2022;4:1256–1264
74 Yoshikawa N, Skreta M, Darvish K, et al. Large language models for chemistry robotics. Auton Robots 2023;47(08):1057–1086
75 Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. ACS Cent Sci 2017;3(05):434–443
76 Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci 2018;4(11):1465–1476
77 Amar Y, Schweidtmann AM, Deutsch P, Cao L, Lapkin A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. Chem Sci (Camb) 2019;10(27): 6697–6706
78 Rinehart NI, Saunthwal RK, Wellauer J, et al. A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C-N couplings. Science 2023;381(6661):965–972
79 Gong Y, Xue D, Chuai G, Yu J, Liu Q. DeepReac+: deep active learning for quantitative modeling of organic chemical reactions. Chem Sci (Camb) 2021;12(43):14459–14472
80 Shields BJ, Stevens J, Li J, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature 2021;590(7844):89–96
81 Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. Nature 2020;583(7815):237–241
82 Wang X, Hsieh CY, Yin X, et al. Generic interpretable reaction condition predictions with open reaction condition datasets and

unsupervised learning of reaction center. Research (Wash D C) 2023;6:0231
83 Andronov M, Voinarovska V, Andronova N, Wand M, Clevert DA, Schmidhuber J. Reagent prediction with a molecular transformer improves reaction data quality. Chem Sci (Camb) 2023;14(12): 3235–3246
84 Maser MR, Cui AY, Ryou S, DeLano TJ, Yue Y, Reisman SE. Multilabel classification models for the prediction of cross-coupling reaction conditions. J Chem Inf Model 2021;61(01): 156–166
85 Schwaller P, Vaucher AC, Laplaza R, et al. Machine intelligence for chemical reaction space. Wiley Interdiscip Rev Comput Mol Sci 2022;12:e1604
86 Schwaller P, Probst D, Vaucher AC, et al. Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 2021;3:144–152
87 Saebi M, Nan B, Herr JE, et al. On the use of real-world datasets for reaction yield prediction. Chem Sci (Camb) 2023;14(19): 4997–5005
88 Probst D, Schwaller P, Reymond JL. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. Digit Discov 2022;1(02):91–97
89 Sandfort F, Strieth-Kalthoff F, Kühnemund M, Beecks C, Glorius F. A structure-based platform for predicting chemical reactivity. Chem 2020;6:1379–1390
90 Schwaller P, Vaucher AC, Laino T, Reymond JL. Prediction of chemical reaction yields using deep learning. Mach Learn Sci Technol 2021;2:015016
91 Yin X, Hsieh CY, Wang X, et al. Enhancing generic reaction yield prediction through reaction condition-based contrastive learning. Research (Wash D C) 2024;7:0292
92 Ma Y, Zhang X, Zhu L, et al. Machine learning and quantum calculation for predicting yield in Cu-catalyzed P-H reactions. Molecules 2023;28(16):5995
93 Maruoka T, Yada A, Sato K, Matsubara S. Machine learning that proposes reaction conditions and yields for wittig-type methylenation of aldehydes with bis(iodozincio)methane in a flow-microreactor. Chem Lett 2023;52:397–399

