Journal of Medicinal Chemistry

pubs.acs.org/jmc

Perspective

Generative Models for De Novo Drug Design

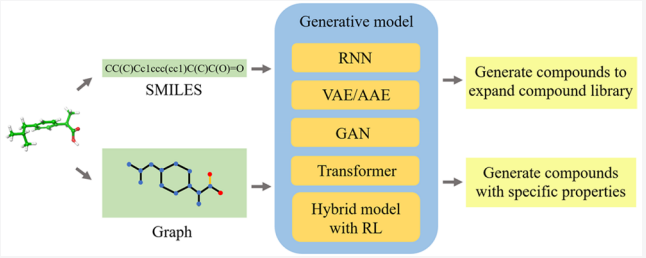
Xiaochu Tong, Xiaohong Liu, Xiaoqin Tan, Xutong Li, Jiaxin Jiang, Zhaoping Xiong, Tingyang Xu, Hualiang Jiang,* Nan Qiao,* and Mingyue Zheng*

Cite This: *J. Med. Chem.* 2021, 64, 14011−14027

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Artificial intelligence (AI) is booming. Among various AI approaches, generative models have received much attention in recent years. Inspired by these successes, researchers are now applying generative model techniques to de novo drug design, which has been considered as the "holy grail" of drug discovery. In this Perspective, we first focus on describing models such as recurrent neural network, autoencoder, generative adversarial network, transformer, and hybrid models with reinforcement learning. Next, we summarize the applications of generative models to drug design, including generating various compounds to expand the compound library and designing compounds with specific properties, and we also list a few publicly available molecular design tools based on generative models which can be used directly to generate molecules. In addition, we also introduce current benchmarks and metrics frequently used for generative models. Finally, we discuss the challenges and prospects of using generative models to aid drug design.



INTRODUCTION

The development of a new drug is a complex process with high cost, high risk, and a long cycle. It takes billions of dollars and 10−15 years for an innovative drug to be developed and finally put on the market.[1] The development of new drugs involves multiple steps, such as discovery and optimization of lead compounds and clinical research, among which the inefficient discovery of early lead compounds is still an important issue that needs to be resolved urgently.

There are currently some open accessible resources of chemical compounds and their biological activities, such as ChEMBL,[2] PubChem,[3] and ChemSpider.[4] The number of compounds of these databases is generally at the level of several million. However, the chemical space of potential drug-like compounds is much larger, with estimates ranging from 10^{23} to 10^{60}.[5] It is therefore extremely challenging how to explore such a huge space more effectively and to find new molecules with special properties.

In the early stage of rational drug design, molecules with novel structures can be constructed by combining fragments of existing compounds[6] or using optimization algorithms such as genetic algorithms.[7−9] With the rapid development of computer science and high-performance computing, artificial intelligence (AI) approaches have been successful in fields such as image processing, pattern recognition, and natural language processing. In recent years, machine learning, especially deep learning, has also been applied to drug discovery, such as predicting compound properties and activities and their interaction with protein targets. In the past few years, deep generative models have attracted increasing attention, which try to learn the probability distribution of the training data, extract representative features, produce a low-dimensional continuous representation, and eventually generate new data by sampling from the learned data distribution. Different applications of generative models have shown extraordinary results in the generation of images,[10] text,[11] speech,[12] and music.[13] The development of generative models has also brought new ideas for solving the difficult problem of drug design and has been considered to be one of the most promising approaches for drug design.[14] When applied to generate molecules, the essence of the generative model is to learn the distribution of molecules in the training set, so as to obtain molecules that are similar to but different from the molecules in the training set. By combining evolutionary algorithm or reinforcement learning, the specified properties of generated molecules can be further optimized. The molecular representation in generative models can be in any form, including chemical fingerprints, simplified molecular input line entry system (SMILES), molecular graph, three-dimensional structures, etc.

In this Perspective, we focus on the application of generative models in de novo drug design. First of all, we briefly introduce

Received: May 22, 2021
Published: September 17, 2021

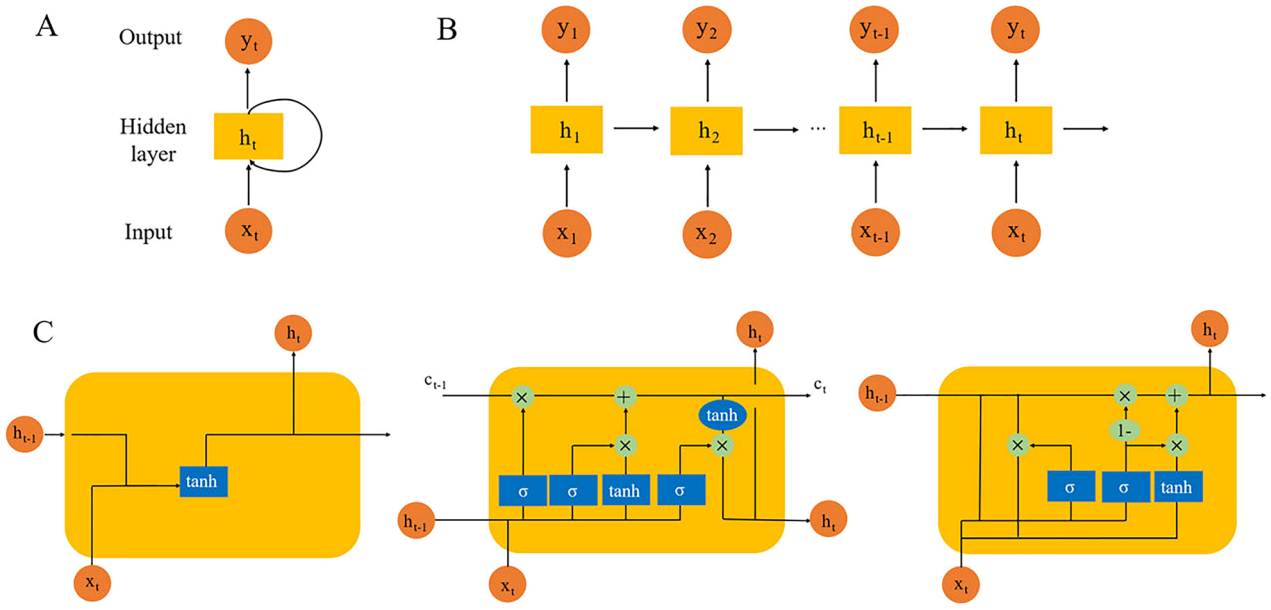


Figure 1. Structure of RNN: (A) the basic network structure of RNN; (B) an unrolled RNN structure; (C) internal structures of basic RNN, LSTM, and GRU.

the frequently used generative models, such as recurrent neural network, autoencoder, generative adversarial network, transformer, and hybrid models combining deep generative models with reinforcement learning. Second, we comprehensively review the latest development in the application of various generative models in drug design and benchmarks and metrics for evaluating their performances. Finally, we discuss the prospect of the generative models for drug design.

PRINCIPLES OF A GENERATIVE MODEL

In this section, generative models are roughly divided into four categories, including the models based on recurrent neural network (RNN), autoencoder (AE), generative adversarial network (GAN), and transformer and hybrid models combining deep generative models with reinforcement learning (RL). The basic principles and recent developments of these popular generative models are described as follows.

RNN-Based Models. The RNN-based model has been used in the field of natural language processing[11,15] and now has also been widely used in other different fields.[16−18] The first study on RNNs was the Hopfield model proposed by Hopfield.[19] However, due to its difficulty in implementation, it has been used less in practice. Generally, the simple recurrent network models proposed by Jordan[20] and Elman[21] are considered to be the basic version of the current RNN.

Figure 1A shows the basic network structure of the RNN, where, through the loop connection on the hidden layer, the current state of the network at the previous time can be received at the current time and the network state at the current time can be further transmitted to the next time. Namely, as an unrolled RNN in Figure 1B, the hidden unit h_t receives data from two aspects at time t, the hidden unit value h_{t-1} at the previous time of the network and the current input data x_t, respectively, and two outputs will be obtained through the calculation of the value of the hidden unit, an output vector y_t and an updated hidden unit h_t.

$$h_t = f(h_{t-1}, x_t) \quad (1)$$

$$y_t = O(h_t) \quad (2)$$

The parameters of the RNN can be estimated by minimizing the cost function. Back propagation through the time algorithm can be used to update the parameters in the network, but it often results in the phenomena of "gradient explosion" and "gradient disappearance" in the RNN model.[22] Subsequently, these problems have been mitigated by using microstructures such as long short-term memory (LSTM) cells[23] and gated recurrent units (GRUs).[24] Their internal structures are more complex and help to store and update information selectively (Figure 1C). Hochreiter et al. proposed an LSTM unit with controlled gates for input, forget, and output.[23] The elaborate "gate" structure was used to remove or enhance the information to the cell state. The LSTM cell uses a more controlled flow of information to determine which information can be retained and which can be discarded. LSTM implements a more refined internal processing unit, which can maintain its internal state to extend the time of sequential input in the RNN, thereby improving the performance of the RNN. Further studies revealed that GRU is a simplified implementation of LSTM architecture and can alleviate the problem of gradient disappearance and explosion at a lower computational cost.[24]

When the RNN model is applied to de novo drug design, the molecules can be represented in the form of sequences, such as by using SMILES. Specifically, after training with a large number of SMILES strings, the RNN model can be used to generate a new valid SMILES that is not included in the original data set, and thus can be considered as a molecular structure generative model.

In addition, due to the needs of generating molecules that meet required conditions, additional information on molecules is incorporated into the RNN models by simply transforming them into the initial state of the network. And the conditional negative log-likelihood (CNLL) of the probability of sequence generation is defined as

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

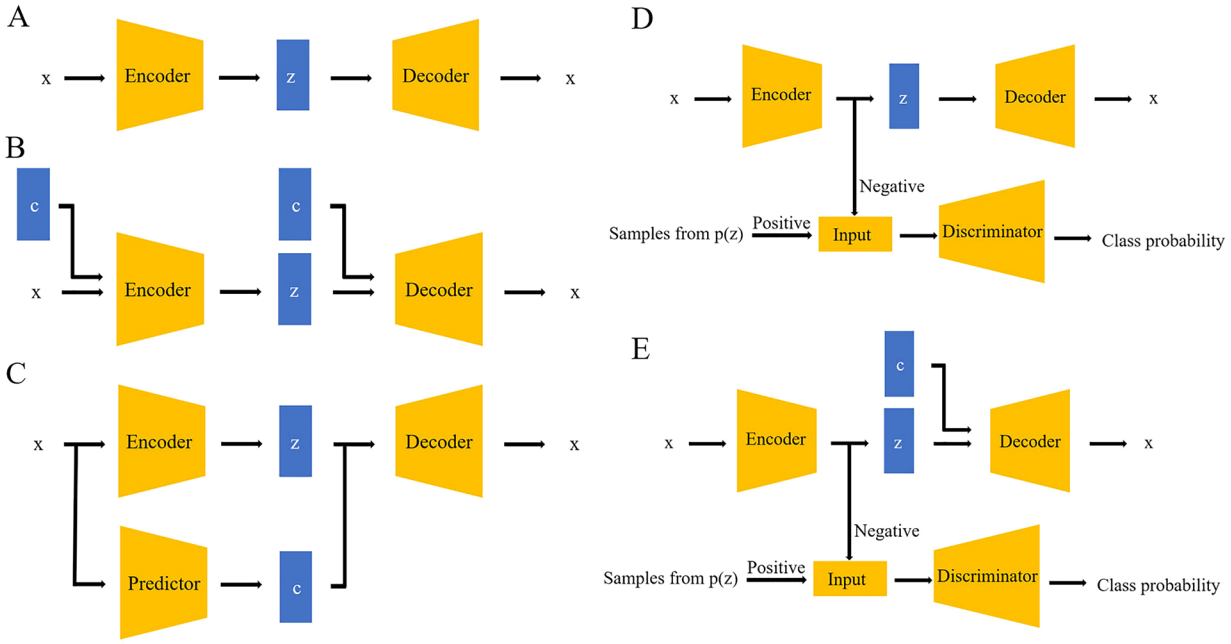


Figure 2. Structure of VAE, AAE, and their corresponding conditional generation models. (A) The structure of VAE. (B) The structure of ContidionalVAE with all labeled molecules. (C) The structure of ContidionalVAE combined with a predictor for unlabeled molecules. (D) The structure of AAE. (E) Simplified version of ConditionalAAE with all labeled molecules.

$$\mathrm{CNLL}(s|c) = -\left[\ln P(x_1 = y_1|c) + \sum_{t=2}^{N} \ln P(x_t = y_t|x_{t-1} = y_{t-1}, \ldots, x_1 = y_1, c)\right] \quad (3)$$

where the condition vector is indicated as c. x_t is the predicted character, y_t is the character in SMILES sequences, and N represents the length of the sequence.

AE-Based Models. Autoencoder[25] is composed of two networks: the encoder maps the high-dimensional data to the low-dimensional representation, and the decoder reconstructs the original input as output given the low-dimensional representation. The autoencoder is trained repeatedly to minimize the deviation between the reconstructed output and the original input, and the goal of the autoencoder is to find a more compact representation of samples. The variational autoencoder (VAE) and adversarial autoencoder (AAE) modify the classical AE with some additional constraints to learn the latent representation from the input data. Different from the aim of AE, these models are designed to learn the probability distribution of the data set, thereby generating samples that are similar to but different from the data set. Figure 2 compares the structures of VAE and AAE. In 2013, Kingma et al. proposed a generative network structure based on variational Bayes inference.[26] Different from the autoencoder, the output of the encoder and decoder in VAE is the probability distribution of the data in the latent and initial space, respectively. In VAE, the continuous representation z is interpreted as a latent variable and $p(z)$ is the prior distribution following a Gaussian distribution. A probabilistic decoder is defined by a likelihood function $p_\theta(x|z)$ with parameters θ, and the encoder approximates the posterior distribution with a model $q_\varphi(z|x)$ paramerized by φ. The goal of the model is to maximize the probability of each x in the training set by formula $p(x) = \int p_\theta(x|z)\, p(z)\, dz$, but this integral is intractable to compute. Therefore, $q_\varphi(z|x)$ is induced as an estimate of posterior distribution $p(z|x)$, and the goal is replaced by maximizing the evidence lower bound as follows,[26] which is always less than or equal to $\log p_\theta(x)$:

$$L(\theta, \varphi; x) = E_{q(z|x)}[\log p_\theta(x, z) - \log q_\varphi(z|x)] \le \log p_\theta(x) \quad (4)$$

Here this formula can be also written as

$$L(\theta, \varphi; x) = E_{q(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}[q_\varphi(z|x), p(z)] \quad (5)$$

According to the above formula, it should maximize the chance of reconstruction $p_\theta(x|z)$ and minimize the Kullback–Leibler (KL) divergence[27] between $q_\varphi(z|x)$ and the prior distribution $p(z)$ to maximize $L(\theta, \varphi; x)$.

The conditional variational autoencoder (ConditionalVAE) applied in de novo drug design is derived from the semi-supervised variational autoencoder (SSVAE) proposed by Kingma et al.[28] Specifically, there are two different scenes of introduction conditions. When the molecular properties considered as conditions can be directly calculated for all molecules, these conditions can be incorporated into the inputs of the encoder and the decoder (Figure 2B). The corresponding objective function is given by eq 6

$$L(\theta, \varphi; x, c) = E_{q(z|x,c)}[\log p_\theta(x|z, c)] - D_{\mathrm{KL}}[q_\varphi(z|x, c), p(z|c)] \quad (6)$$

where c is a condition vector.

In the other scene, if the conditions cannot directly label all molecules, like biological activity against specific targets, VAE should combine with a predictor network to predict properties for those unlabeled molecules, and the condition vector c is considered as a latent variable from the predictor (Figure 2C). The objective function for unlabeled molecules is shown as follows:

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

$$L(\theta, \varphi; x) = E_{q(c,z|x)}[\log p_\theta(x|z, c)] - D_{KL}[q_\varphi(c|x), p(c)] - E_{q(c|x)}[D_{KL}(q_\varphi(z|x, c), p(z))] \quad (7)$$

In 2015, Alireza Makhzani et al. proposed adversarial autoencoder (AAE).[29] AAE (Figure 2D) is similar to VAE, but its characteristic is to add a discriminant neural network in the architecture, which is derived from the GAN model.[30] AAE uses adversarial training with a discriminator D, which can distinguish the generator's latent distribution from the prior to avoid the use of KL divergence. The encoder of the model can be regarded as a generator G, and the output of $G(x)$ fools the discriminator D by mimicking the prior arbitrary distribution $p(z)$. Meanwhile, the discriminator D is trained to discriminate between the latent distribution from the encoder and the prior $p(z)$. The model is optimized by the following formula:

$$\min_{G,\theta} \max_{D} E_{x\sim p_{data}(x)}[\log D(G(x))] + E_{z\sim p(z)}[\log(1 - D(z))] - E_{x\sim p_{data}(x)}[\log p_\theta(x|G(x))] \quad (8)$$

Conditional extension of AAE was also mentioned by Makhzani et al.,[29] including supervised AAE and semisupervised AAE. For supervised AAE, the decoder reconstructs molecules from latent vectors and condition vectors (Figure 2E). The condition of unlabeled molecules should be generated in semisupervised AAE, so an additional adversarial network is imposed to ensure the posterior distribution of c matches the predefined categorical distribution.

GAN-Based Models. The concept of a generative adversarial network was first proposed in 2014 by Goodfellow, inspired by the game theory of two-person zero-sum game.[30] The generative adversarial model includes a generator G and a discriminator D (Figure 3A). Generally, the generator learns to map the random noise to the specific distribution that needs to be close to the data distribution, while the discriminator determines whether the input is real data or the generated sample from the generator which is usually a binary classifier. Once the model is well-trained, new samples can be obtained from the generator. Specifically, in the adversarial process, two neural network models, generator G and discriminator D, are trained at the same time, so that D can find the hidden pattern in the input data to accurately distinguish the real data from the data generated by G, and G will iterate through optimizing the weights for matrix multiplication of data sampling to learn to deceive the well-trained D. Overall, the essence of the GAN model is the zero-sum game where D and G compete with each other. The following shows the objective function in the original paper on GAN:

$$\min_{G} \max_{D} E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p(z)}[\log(1 - D(G(z)))] \quad (9)$$

In the above formula, $p_{data}(x)$ is the real data distribution and $p(z)$ is a prior probability distribution. The discriminator D is trained to maximize the probability of assigning the correct label to both training examples and samples from G, and the generator G is simultaneously trained to minimize $\log(1 - D(G(z)))$.

Conditional generative adversarial network (Conditional-GAN)[31] is a variant of GAN, which is conditioned by adding extra information c into both the generator and the discriminator (Figure 3B). Condition vector c and input noise z are fed into the generator, and in the discriminator, condition vectors c concatenated with training samples are used as inputs. The objective function is presented as eq 10, in which the representations are the same as eq 9 except for the condition vector c:

$$\min_{G} \max_{D} E_{x\sim p_{data}(x)}[\log D(x|c)] + E_{z\sim p(z)}[\log(1 - D(G(z|c)))] \quad (10)$$

Transformer Models. Transformer is a new model that was proposed recently, showing state-of-the-art performance in natural language processing.[32,33] The original version of transformer consists of encoder and decoder (Figure 4), and the key of this model is the attention mechanism, which can consider long-range dependencies in sequences. In detail, there are three vectors including the key, query, and value vectors, and the corresponding attention is represented as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

where d_k is the dimension of key and query vectors and used to scale the dot products of these vectors.

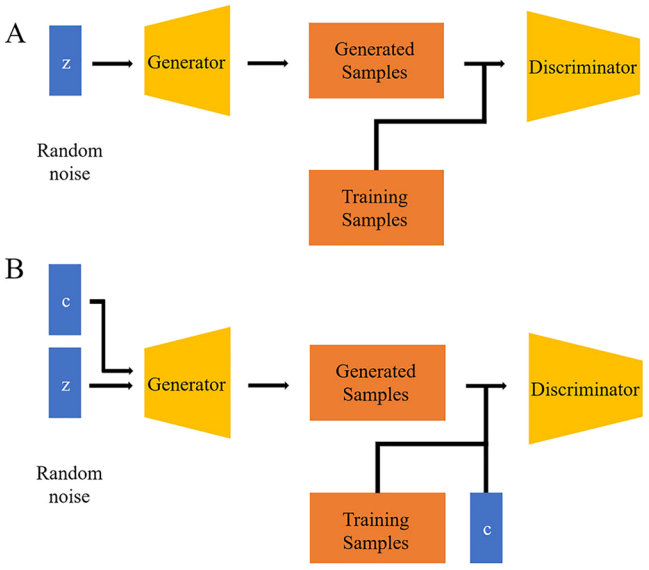


Figure 3. Structure of GAN (A) and ConditionalGAN (B).

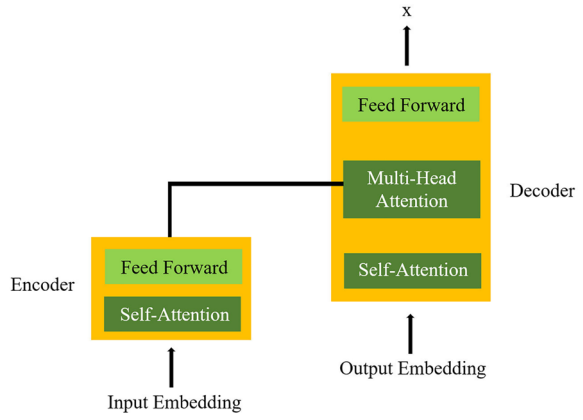


Figure 4. Simplified structure of the transformer.

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

Table 1. A List of Some Generative Models for Molecular Design

model	data set	code	references
BIMODAL	ChEMBL	https://github.com/ETHmodlab/BIMODAL	43
REINVENT	ChEMBL; ExCAPE-DB	https://github.com/MarcusOlivecrona/REINVENT	36
ChemicalVAE	QM9; ZINC	https://github.com/aspuru-guzik-group/chemical_vae	41
GrammarVAE	ZINC	https://github.com/mkusner/grammarVAE	44
SD-VAE	ZINC	https://github.com/Hanjun-Dai/sdvae	45
ORGAN	ZINC; GDB-17	http://github.com/gablg1/ORGAN	38
ORGANIC	ZINC; GDB-17; Harvard Clean Energy Project	https://github.com/aspuru-guzik-group/ORGANIC	42
LatentGAN	ChEMBL; ExCAPE-DB	https://github.com/Dierme/latent-gan	46
ARAE	QM9; ZINC	https://github.com/gicsaw/ARAE_SMILES	47
Onco-AAE	NCI-60 cell line assay data	https://github.com/spoilt333/onco-aae	48
LigGPT	MOSES data set; GuacaMol data set	https://github.com/devalab/liggpt	49
molecule_structure_generation	BindingDB	https://github.com/dariagrechishnikova/molecule_structure_generation	50
MolRNN	ChEMBL	https://github.com/kevinid/molecule_generator	51
CGVAE	QM9; ZINC; CEPDB	https://github.com/Microsoft/constrained-graph-variational-autoencoder	52
MolGAN	QM9	https://github.com/nicola-decao/MolGAN	53
GCPN	ZINC	https://github.com/bowenliu16/rl_graph_generation	54
NeVAE	QM9; ZINC	https://github.com/Networks-Learning/nevae	55
GENTRL	ZINC; ChEMBL; Integrity	https://github.com/insilicomedicine/gentrl	56
JT-VAE	ZINC	https://github.com/wengong-jin/icml18-jtnn	57
DeLinker	ZINC; CASF	https://github.com/oxpig/DeLinker	58
DL4chem	QM9; COD; CSD	https://github.com/nyu-dl/dl4chem-geometry	59
GRAPHDG	ISO17 data set	https://github.com/gncs/graphdg	60
MOLGYM	QM9	https://github.com/gncs/molgym	61

Considering that the order of tokens in sequences is not contained in the attention mechanism, additional position information is injected into the inputs. Specifically, sine and cosine functions are used in the form of the following formulas

$$PE_{(\mathrm{pos},2i)} = \sin\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right) \tag{12}$$

$$PE_{(\mathrm{pos},2i+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right) \tag{13}$$

where pos represents the position, i represents the dimension, and d_{model} is the size of embedding.

Hybrid Models. Hybrid models combining deep generative models with reinforcement learning[34,35] have been applied to generate de novo molecules biased to the desired properties.[36−38] Reinforcement learning is a goal-oriented machine learning method that uses environmental feedback as input and adapts to the environment. Its main idea is to interact with the environment with trial and error to find the optimal behavior strategy which mimics the basic way for humans or animals to learn.[35] The core principle of reinforcement learning is to learn a series of actions that will guide the agent to achieve its goal or maximize its objective function. If a certain action of the agent leads to a positive reward from the environment, that is, a strengthening signal, then the trend of each subsequent action of the agent will be strengthened. Otherwise, the agent's tendency to produce this action is weakened. This is consistent with the principle of conditioned reflex in physiology.

APPLICATIONS OF GENERATIVE MODELS IN DRUG DESIGN

In this section, we review the latest developments in the application of various generative models in de novo drug design, which are mainly used to expand existing compound libraries for virtual screening and generate compounds with specific properties. These include not only in silico applications of generative model algorithms to design and optimize molecules but also real-world cases with experimental verification results. Finally, we also summarize a few publicly available de novo molecular design tools based on generative models for users with less experience in coding or the knowledge of artificial intelligence.

Generating Compounds and Expanding Compound Libraries. With the development of generative models and their application in the field of chemistry, our capability to explore unknown chemical spaces can be enhanced. For example, Josep Arús-Pous et al. have shown that a RNN model trained with 0.1% of a database reproduced 68.9% of the entire database after training.[39] Moreover, a generative model can be trained to determine a joint probability distribution $p(x, y)$, i.e., the probability of observing both a molecular representation (x) and its physical property (y). It is thus possible to perform inverse design $p(x|y)$ to design new compounds with specific properties, by conditioning the probability on a property (y).[40] Simple molecular representations such as SMILES and molecular fingerprints have been frequently used as the input to the model. Although they are concise and convenient, both of them have their own limitations. Fingerprints are fixed-length representations that must be extremely large to encode all possible substructures without overlap. For SMILES-based representations, the molecular structure is encoded according to certain syntax rules. Generated SMILES that do not meet these rules will be considered as invalid molecules, and additional checks are needed to remove these invalid SMILES. Recently, many efforts have been devoted to the molecular graph that directly represents the molecule structures. Meanwhile, by using fine-tuning, Bayesian optimization,[41] transfer learning,[16] and

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

reinforcement learning[36,37,42] on the model, the new generated molecules can be optimized, and the generative model can produce molecules with desired properties. Table 1 shows generative models for generating compounds and expanding the compound library.

Using SMILES or Molecular Fingerprints as a Representation. Many deep generative model techniques have been developed specifically for sequence generation. Therefore, when generative models are applied to de novo drug design, SMILES, which encodes a molecular graph compactly as a line of text, has been first used as input to the generative model to generate new molecules. SMILES was developed by Weininger in the late 1980s,[62] which is a formal grammar that explicitly describes molecular structures using ASCII strings. For example, O for oxygen, c and C for aromatic and aliphatic carbon atoms, and −, =, and # for single, double, and triple bonds. An important feature of SMILES is that it is easy to learn and human-readable compared to most other methods of molecular representation.

Most existing studies on the generative model for generating new molecules have used SMILES as a molecular representation, including RNN, VAE, GAN, and so on. Bjerrum et al. applied an RNN model through training on existing molecules from ZINC compounds encoded as SMILES to generate virtual compound libraries. The properties of the produced molecules are similar to those in training databases.[18] Arús-Pous et al. trained RNN models with a subset of the enumerated database GDB-13 to explore the GDB-13 chemical space. As a result, the trained model reproduced 68.9% of the entire database by using only 0.1% molecule structures of the database.[39]

Segler et al. trained a long short-term memory RNN model based on a large set of molecules from ChEMBL represented by canonicalized SMILES called CharRNN to generate diverse and chemically reasonable molecules. Furthermore, the model produced specifically targeted molecules by performing transfer learning on small sets of molecules with known actives.[16] Moret et al.[63] comprehensively explored the effects of data augmentation, temperature sampling, and transfer learning on the application of generating molecules in low data regimes. First, it is clear that the quality of the generated molecules will not increase indefinitely with the increase of data and sampling temperature, and the combination of these two has been investigated. Meanwhile, the physicochemical properties of a small set of molecules can be captured by transfer learning, and the generative model fine-tuned by molecules with various structures can generate a broad range of structurally diverse molecules.[63] In terms of application, several studies have applied this transfer learning strategy to generate actives for specific targets in practical drug design projects and identified novel candidate compounds that show biological activity against specific targets in vitro and in vivo. Yang et al.[64] reported using the LSTM-based neural network model[16] to train 200,000 compounds from the ChEMBL database and then fine-tuned the model with a data set containing 135 published p300 inhibitors and 576 macrocycle molecules to generate novel p300/CBP inhibitors. As a result, the model generated a focused library containing 672 chemical structures, from which some compounds were selected for synthesis. After further systematic optimization, a set of highly effective inhibitors was obtained. Among them, a potential candidate B026 showed high inhibitory activity against p300/CBP histone acetyltransferases and significant tumor growth inhibition in animal models of human cancer, which has been identified for further preclinical development.[64] Li et al.[65] applied a RNN-based generative model to discover potential inhibitors for Moloney murine leukemia virus kinase 1 (Pim1) and cyclin-dependent kinase 4 (CDK4). They trained the model on the set of randomized SMILES sequences of CDK4 inhibitors and Pim1 inhibitors, and three molecules were selected based on synthetic accessibility. These three molecules contain some difficult-to-attach fragments and thus were further simplified prior to synthesis, leading to MJ-4, MJ-115, and MJ-1055. These molecules verified the inhibitory activity on Pim1 and CDK4. Among them, MJ-1055 had potent inhibitory activity on Pim1 with a IC50 value of 9.6 nM, and it was found to be different from similar molecules protected in the relevant Markush patent. In contrast, MJ-4 showed weak inhibitory activity on CDK4, and MJ-115 also showed significantly reduced activity as compared to known inhibitors with similar structures. Overall, these results well support the applicability and potential of the RNN-based generative model in practical tasks and also suggested that molecules generated by the RNN-based model alone may not maintain the desired activity.[65] Recently, Tan et al.[66] combined the RNN model with a multitask deep neural network to automatically design and optimize antipsychotic drugs targeting multiple G-protein coupled receptors (GPCRs). A multitask neural network was first established to predict the activity of compounds targeting the D_2/5-HT_{1A}/5-HT_{2A} receptor. The generated molecules with high predictive activity were used to expand the fine-tuning set at each iteration during the transfer learning process, which could avoid the newly generated molecules being too similar to those in the training set. A hit compound was obtained through the above process, and it was latter synthesized and evaluated, showing potent activities against D_2, 5-HT_{1A}, and 5-HT_{2A} receptors in biochemical experiments. Furthermore, 10 analogues of hit compounds were designed by introducing different linkers or heterocyclic moieties, and six analogues were selected by the above activity prediction model and evaluated experimentally. Among them, one compound exhibited not only promising activities on the D_2, 5-HT_{1A}, and 5-HT_{2A} receptors in vitro but also a potent antipsychotic effect in animal models, showing good potential for subsequent development.[66] Schneider's group also designed novel and biologically active compounds from scratch by using generative models.[67,68] They developed a generic RNN model that learned the constitution of drug-like molecules from a large set of known bioactive compounds and used 25 fatty acid mimetics known to have agonistic activity on retinoid X receptors (RXRs) and/or peroxisome proliferator-activated receptors (PPARs) to fine-tune the model. By employing the target prediction method SPiDER,[69] molecular shape, and partial charge descriptors to rank the generated molecules, 49 high-scoring compounds were finally obtained. In the end, they selected five compounds for synthesis and experimentally verified their agonistic effects on nuclear receptors, and they found that four of the compounds had nanomolar to low-micromolar receptor modulatory activity in cell-based assays.[67] In another work, they used a similar workflow to design natural-product-inspired retinoid X receptor modulators by using natural products that activate RXR to perform transfer learning. In practice, they fine-tuned the generative model using six natural products that activate RXR and obtained 201 compounds that are suitable for synthesis and predicted to be RXR agonists by SPiDER.[69] WHALES descriptors were used to rank these molecules, and the top 50 compounds were further selected for visual inspection. As a result, four generated

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

compounds were selected for synthesis and in vitro experimental verification and two compounds were confirmed to have potential RXR agonistic activity.[68]

In addition to the conventional forward RNN model, the bidirectional RNN model has also been explored as a new method for SMILES string generation. The bidirectional RNN model BIMODAL used SMILES with a randomly placed starting token during training and showed improved "novelty" value than the forward RNN model.[43] Olivecrona et al.[36] proposed the model called REINVENT which also used an RNN for molecular de novo design, and they introduced a reinforcement learning method to fine-tune the pre-trained RNN, so the model could generate structures with desirable properties, such as molecules that do not contain sulfur, analogues of the drug Celecoxib, and active compounds of dopamine receptor type 2 (DRD2). Specifically, they used this model to produce DRD2 actives and found more than 95% of the molecules produced by this model are predicted to be active, including those that have been experimentally confirmed to be active but not included in the generation model or activity prediction model.[36] Recently, Yoshimori et al.[70] used the generative model REINVENT with a pharmacophore model for the design of discoidin domain receptor 1 (DDR1) inhibitors. They sampled SMILES from the well-trained model and further performed filtering of the generated molecules by pharmacophore scores and binding affinity scores. Subsequently, nine compounds were selected for synthesis, and two compounds were modified during visual inspection. In the end, these synthesized compounds were evaluated for their inhibitory activity against DDR1 and three of them were found to have submicromolar inhibitory activity.[70]

For the VAE and AAE models, the new and valid SMILES strings can be obtained by the decoder to achieve de novo molecular generation. Gomez-Bombarelli et al.[41] first applied a VAE framework[26] to chemical design called ChemicalVAE. In the autoencoder model, they converted molecules represented as SMILES strings into the latent space that captures characteristic features of the training data. The autoencoder can be trained jointly with a property prediction task to optimize molecular properties; however, one of the major problems with this model is that sometimes the produced molecules are invalid or contain undesirable moieties, such as acid chlorides, cyclobutadienes, and so on.[41] In order to solve the above problems, Kusner et al.[44] induced context free grammar (CFG)[71] to give the VAE explicit knowledge about how to produce valid molecules. They proposed grammar variational autoencoder (GrammarVAE) which showed a higher percentage of valid molecules by generating syntactically valid SMILES and also obtained a smoother latent space.[44] However, there are also non-context-free aspects of SMILES strings that went unmodeled; for example, the syntax of SMILES requires that the rings generated must be closed in molecule generation. Therefore, syntax-direct variational autoencoder (SD-VAE) introduced attribute grammar to enrich the CFG with "semantic meaning", and the percentage of valid SMILES generated by training this model was further increased.[45] Blaschke et al.[72] constructed four different AE models including VAE with and without the teacher forcing method, and AAE using a Gaussian or a uniform distribution, and compared the capabilities of these four models in generating compounds. As a result, the teacher-forcing-based VAE model had a much higher fraction of valid SMILES than VAE without teacher forcing, and AAE imposing a uniform distribution showed the largest fraction of valid SMILES and generated the smoothest latent chemical space representation. Furthermore, they trained VAE with Bayesian optimization to generate novel compounds with predicted activity against dopamine receptor type 2.[72]

As a special generative model, GAN has also been applied to molecular generation based on SMILES. One of the first successful applications of GAN in molecule generation was the objective-reinforced generative adversarial network (ORGAN)[38] and its improved version, objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC).[42] Guimaraes et al.[38] presented ORGAN, a GAN framework with RL based on SeqGAN[73] which can optimize the properties of the generated molecules. To improve the stability of the adversarial training, they also implemented a variant ORGAN that utilized the Wasserstein distance[74] and found the generated molecules showed better diversity. Overall, these models can generate molecules that learn the original data distribution, show improvement in the desired metrics, and also maintain diversity of the samples.[38] ORGANIC[42] is an implementation of ORGAN in the direction of chemistry. As described, the main shortcoming of ORGANIC is a large number of invalid molecules and there might be many repetitions in the valid molecules. This can be caused by the roughness of the chemical space, and small changes in chemical space can have dramatic effects on molecular structure.[38] Putin et al. came up with models combining GAN and RL called reinforced adversarial neural computer (RANC)[75] and adversarial threshold neural computer (ATNC).[76] These models are deep neural network architectures based on the ORGANIC paradigm but include differentiable neural computer (DNC),[77] a type of RNN with external memory, as a generator. The DNC controller helps to balance the generator and discriminator during adversarial training, so the models do not suffer from the perfect discriminator problem. In addition, ATNC uses the adversarial threshold (AT) block to act as a filter between the agent and the environment, and it selects the molecules that most match to the training data. As a result, these models showed better performance when compared with the ORGANIC model and can produce valid and unique SMILES strings more stably. Prykhodko et al.[78] combined autoencoder with generative adversarial neural network to produce LatentGAN for de novo molecular design. In this model, SMILES of molecules were not used in GAN directly but first transformed into latent vectors through a heterencoder strategy.[46] This process alleviated the complexity caused by molecules with similar structures that may have different canonical SMILES and reduce the overfitting problem caused by multiple representations of the same molecule.[78] In addition to the combination of AE and GAN, the combination of VAE and GAN was newly proposed, as these two methods are complementary to each other. A model combining these two schemes has two merits. First, it can avoid the insufficiently flexible approximation of posterior distribution in VAE, which may cause unnatural molecules or even invalid outputs. Second, it can avoid the difficulty in handling discrete variables in GAN, which may cause a low-diversity problem and a repeated generation of molecules. Hong et al.[47] proposed to use the framework of adversarially regularized autoencoder (ARAE),[79] which takes advantage of GAN- and VAE-based models, to generate valid and unique molecules.[47] Specifically, this model is based on VAE, but the distribution of latent variables in this model is not approximated by predefined functions but obtained through an adversarial training process like GAN. Meanwhile, in

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

order to avoid the difficulty of discrete data processing, continuous latent variables are used in adversarial training instead of discrete molecular structures.

Transformer[33] is an advanced model architecture for solving some problems of RNN-based models, which has been used for de novo drug design. LigGPT[49] is a mini version of the generative pre-training transformer (GPT) model[80] for molecular generation, which can learn long-range dependencies in SMILES, like ring closures. The masked self-attention mechanism was applied in the model because, when predicting the next token in a sequence, only the attention in front of this token should be utilized. Mandhana et al. also constructed the model Transformer-XL[81] based on SMILES to generate molecules, and this model successfully considered variable-length molecular sequences. Grechishnikova[50] proposed a de novo drug generative model based on the Transformer architecture[33] which considered molecular generation as a translational task from protein sequences and SMILES of molecules. The goal of this model is to generate a lead compound for a specific protein with only sequence information.[50]

In addition to using SMILES to represent molecule structure as input of the generation model, other molecular representations have also been tried. Gomez-Bombarelli et al.[41] tested InChI[82] as a representation of molecule input into the VAE model to generate a new molecule and found that its performance is worse than SMILES, possibly due to the syntax involved in counting and the arithmetic being more complex.[41] Some examples have tried molecular fingerprints as molecular representations, but the limitation of generating fingerprints as output is that they cannot be directly converted into real molecular structures. One alternative approach is to use the generated fingerprints to select compounds from the compound library based on molecular fingerprints similarity, but this approach can only be used for screening rather than designing from scratch. Kadurin et al.[48] first proposed the application of AAE to generate novel compounds for cancer treatment. Specifically, they converted the SMILES string provided by PubChem into 166-bit molecular access system (MACCS) chemical fingerprints,[83] used a vector of binary fingerprints and inhibition concentration of the molecule as the model input and output, and trained the model with NCI-60 cell line assay data of 6252 compounds profiled on the MCF-7 cell line. This model was trained to encode and reconstruct not only molecular fingerprints but also experimental concentration. Finally, the output was applied to screen 72 million compounds on PubChem to find candidate molecules with anticancer properties. As a result, 69 compounds belonging to various chemical classes were selected, and it was found that some of them had already been used as anticancer agents for the treatment of various cancer types.[48] Recently, they proposed an adapted AAE model called drug generative adversarial network (druGAN), which could be trained with much larger molecule data sets. Compared with the VAE model, this model showed better capacity and efficiency in generating new molecules with specific anticancer properties.[84] Bian et al.[85] came up with a deep convolutional generative adversarial network (dcGAN) model to screen and design new compounds for cannabinoid (CB) receptors. In this model, the convolutional neural network (CNN) model was used as a discriminator, while the reverse convolution process was used as a generator. In order to determine the appropriate architecture and input data structure of CNN involved, they explored various combinations of network architectures and molecular fingerprints. Finally, the discriminator was established based on the LeNet-5 architecture, and AtomPair fingerprints were selected as the input of the dcGAN model to represent small molecules.[85]

Using a Molecular Graph as a Representation. Although most of the previous molecular generation models use SMILES to represent molecules, SMILES has its own limitations. For example, the literal meaning of the SMILES string is different from the molecular structure it represents, so it may not be appropriate to directly apply existing natural language processing models. Moreover, producing valid SMILES strings requires the model to learn semantic rules that are not relevant to the molecular structure, such as SMILES syntax and atomic ordering, which adds an unnecessary burden to the training process and needs additional checks to remove invalid SMILES after sampling. With the development of graph neural network (GNN), it is a feasible choice to directly use a graph to represent molecules and to operate in the graph space to generate molecules. Chemical checks can be performed, such as valency checks on the molecular graph directly; at the same time, the partially generated molecular graphs can be interpreted as substructures, which helps us make full use of all of the generated molecules. However, the design of the deep generation models for graphs is not easy to implement, as we need to deal with the problems of discrete structure, permutation invariant, and variable size.

Simonovsky and Komodakis[86] trained GraphVAE formulated in the framework of VAE[26] to generate molecular graphs. This model was defined to generate probabilistic fully connected graphs of a predefined maximum size and aligned the generated graphs to the ground truth by using a standard graph matching algorithm. One potential limitation of this model is it can be only applied to generating small graphs.[86] Li et al.[87] introduced a graph generation model similar to Johnson's work,[88] which used a sequential process for the graphs, generated one node at a time, and created edges one by one to connect each node to the existing partial graph. In this way, the graph was converted into a structure to build the sequence of actions. The author modeled this sequential decision process by using a graph network. This graph model was trained to generate molecules with less than 20 heavy atoms based on the ChEMBL data set, which generally outperformed the LSTM model on the same graph generating sequences.[87] Later, Li et al.[51] proposed sequence graph generators MolMP and MolRNN. The graph generation of the former architecture model was taken as a Markov process, and the latter model MolRNN depended both on the current state of the graph and the history. In these models, the training set covers larger compounds containing 50 heavy atoms in the ChEMBL data set.[51] Liu et al.[52] put forward a sequential generative model based on VAE architecture called CGVAE. They introduced gated-graph neural networks[89] in the encoder and decoder of the model and employed valency masks to enforce chemical rules for generating molecules to ensure that new generated molecules are always valid. Unlike Li's work[87] mentioned above, this model is conditioned on the current partial graph rather than on a full history of the generation sequence to avoid overfitting problems.[52] Cao and Kipf[53] adapted GAN to construct the model called MolGAN which operated directly on graph-structured data. It is an implicit, likelihood-free generative model that uses graph convolution and a node aggregation operator to obtain a permutation-invariant discriminator. This model can produce a high proportion of valid and novel compounds while having a low

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

score in uniqueness due to the mode collapse, which is the main failure of GAN architecture.[53] You et al.[54] generated molecule graphs based on the graph convolutional policy network (GCPN) model. This model combined graph representation, reinforcement learning, and adversarial training in a unified framework, enabling the generation of valid molecular graphs. Furthermore, this method could directly optimize the properties of the molecular graph to generate goal-directed molecules. The results showed that, when compared with advanced methods such as JT-VAE[57] and ORGAN,[38] GCPN was superior in molecular property optimization and property targeting.[54]

Samanta et al.[90] proposed a variational autoencoder for graphs where the encoder and decoder were specially designed. The probabilistic encoder learned to aggregate information from a different number of hops away from a given node and then mapped this aggregate information into a continuous latent space, which could encode graphs with a variable number of atoms. Moreover, the decoder jointly represented all edges as an unnormalized log probability vector, which was then fed a single edge distribution, and this allowed for an efficient inference algorithm and decoding. The result showed that the trained autoencoder can find a smooth latent representation of molecules and generate new molecules with higher "validity" and "novelty".[90] Later, Samanta et al.[55] further improved this model and proposed NeVAE. In the decoder of this model, the spatial coordinates of the atoms of the generated molecules can be provided. Furthermore, a gradient-based algorithm was developed to optimize the decoder so that it learns to generate molecules that maximize the value of certain properties. Experiments have shown that, compared with other graph-based models, this model could find reasonable, diverse, and novel molecules more effectively. Besides, this model can also help to discover molecules with low potential energy values by optimizing the spatial configuration of molecules.[55]

Generative tensorial reinforcement learning (GENTRL) is a deep generative model for de novo small molecule design proposed by Zhavoronkov et al.[56] The model specifically combined the algorithms of reinforcement learning, variational inference,[91] and tensor decomposition,[92] and three different self-organizing maps[93] were used as reward functions. Recently, GENTRL was successfully used to find potent inhibitors of DDR1, a kinase target associated with fibrosis and other diseases. Six lead candidates were identified from the new generated compounds, one of which showed good efficacy and pharmacokinetic properties in mice. In this work, effective inhibitors of DDR1 were discovered in 21 days using GENTRL, and the design, synthesis, and experimental testing were completed in a total of 46 days, demonstrating the potential of this method for rapid and effective molecular design.[56]

Different from generating molecular graphs atom by atom, using valid chemical substructures as nodes in the graph is considered a promising method. JT-VAE generates molecular graphs in two phases: first, it generates a tree-structured scaffold which contains subgraph components extracted from the training set, and then, the subgraphs are combined into a molecule by a graph message passing network. In this way, the "validity" of generated molecules can be further improved.[57] Imrie et al.[58] developed a graph generation model called DeLinker. This model took two fragments or partial structures as input and simultaneously combined their three-dimensional structure information, including the distance between the fragments and their relative directions, to generate a molecule containing these two substructures. The results showed that this method can be applied to fragment linking, scaffold hopping, and proteolysis targeting chimera design.[58]

In addition to generating new molecular graphs, there are studies generating three-dimensional molecular structures, which are important in designing molecules with high biological activity. Mansimov et al.[59] proposed a conditional deep generative graph neural network DL4Chem that generates corresponding molecular conformations from molecular graphs by learning the energy function. In the model, conditional variational autoencoder[26] was used for the construction stage of the energy function, graphs were modeled using a messaging neural network, and three-dimensional coordinate vectors of all atoms were used to represent molecular conformations.[94] Compared with the traditional force field methods, the conformations generated by this model were closer to the ground-truth conformation on average with much reduced calculation cost. This method could provide the initial coordinates for the conventional force-field-based methods.[59] Different from the above method, graph distance geometry (GRAPHDG) proposed by Simm et al.[60] used pairwise Euclidean distances between atoms to describe molecule conformation, including edges between the bonded atoms and auxiliary edges between second and third neighbors, which is invariable to rotation and translation.[60] In the framework of conditional variational autoencoder,[26] this generative model can obtain the corresponding set of atomic distances according to molecular graph. By combining an Euclidean distance geometry (EDG) algorithm,[95] the molecular conformation can be obtained. As a result, compared with the classical EDG method[96] in RDKit and the above machine learning method DL4Chem,[59] GRAPHDG showed state-of-the-art performance in the new benchmark they established. Recently, Li et al.[97] built a new generative model architecture named L-net, which can directly generate drug-like molecules with topological and 3D structures. The model was trained using drug-like molecules from the ChEMBL data set, with their 3D coordinates calculated by RDKit. As a result, the model can generate chemically correct, conformationally valid, and drug-like molecules. Furthermore, they combined L-Net with a reinforcement learning method Monte Carlo tree search (MCTS) to optimize molecules targeting tyrosine-protein kinase ABL1. Specifically, they used the functional group that exerts inhibition activity on the known active molecule asciminib as the seed structure and optimized the rest of the structure to obtain higher binding affinity. The result showed that this model can generate molecules with high predicted binding affinity, and the generated molecules had similar interaction modes and predicted binding affinity as compared to known inhibitors.[97] Simm et al. proposed a novel RL formulation by using quantum mechanics to guide molecular design,[61] where the reward function is based on the electronic energy and is approximated by the semiempirical Parametrized Method 6[98] in SPARROW.[99,100] Considering that the properties of molecules are invariant under translation and rotation, the internal coordinates of atoms with respect to existing atoms, like the distance, angle, and dihedral angle, are learned by the agent first, and then, these internal coordinates are mapped to Cartesian coordinates. These procedures allow us to calculate quantum-mechanical properties directly. In the molecular generation process, the agent tries to take atoms from a given bag and place them on a 3D canvas.[61] This sequential generation of atoms in Cartesian coordinates to obtain molecules expands the class of molecules that can be generated and allows the generation of systems consisting of multiple molecules.

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

Currently, this model is limited to designing molecules with known molecular formulas and further exploration is needed to increase its scalability.

Conditional Molecular Design. Most molecular design tasks require generating compounds that meet specific requirements. In addition to optimizing the generated new molecules by using methods like fine-tuning, transfer learning, and reinforcement learning, many efforts have been made to modify the previous generative model, to establish conditional generative models. These kind of models directly incorporate the information on molecular properties coupled with molecular structure information, which can guide molecular generation to specific areas of the chemical space associated specific conditions. Therefore, conditional molecular design samples new molecules from a conditional generative distribution without any additional optimization process. Besides, the conditional models can be easily adapted to consider multiple target properties simultaneously. Some generative models for conditional molecular design are summarized in Table 2.

Table 2. A List of Some Generative Models for Conditional Molecular Design

model	data set	code	references
CVAE	ZINC	https://github.com/jaechanglim/CVAE	101
SSVAE	ZINC	https://github.com/nyu-dl/conditional-molecular-design-ssvae	102
CARAE	ZINC	https://github.com/gicsaw/ARAE_SMILES	47
cRNN	ChEMBL; ExCAPE-DB	https://github.com/pcko1/Deep-Drug-Coder	103
LigGPT	MOSES data set; GuacaMol data set	https://github.com/devalab/liggpt	49
conditional MolRNN	ChEMBL	https://github.com/kevinid/molecule_generator	51

Lim et al.[101] presented a molecular generative model based on conditional variational autoencoder,[28] which could impose certain conditions on latent space. Specifically, the model may generate desired drug-like molecules by controlling five properties, i.e., molecular weight (MW), LogP, number of hydrogen bond donors (HBDs), number of hydrogen bond acceptors (HBAs), and total polar surface area (TPSA), simultaneously. During training, these target properties were formed as a predefined condition vector and concatenated with a latent vector. It was possible to adjust LogP without changing others and generate molecules with a certain property beyond the range of the training set. However, this model showed a low success rate of generating a desirable molecule, which was possibly caused by strong correlation among the properties.[101] Kang and Cho[102] built a model to conditionally generate molecules using the regression version of semisupervised variational autoencoder (SSVAE).[28,102] Instead of defining a condition vector in advance, they used a property prediction model to generate continuous-valued properties as given target condition; thus, SSVAE efficiently generates new molecules that satisfy the target condition without any additional optimization process. This model can take full advantage of a small portion of labeled molecules to improve the performance, which may reduce the cost of labeling molecules.[102] Polykovskiy et al.[104] improved supervised adversarial autoencoders proposed by Makhzani et al.[29] with several disentanglement approaches, called entangled conditional adversarial autoencoder (ECAAE). In this work, they applied the AAE model to generate an inhibitor of Janus kinase 3 (JAK3) and found a promising hit compound which showed good in vitro activity and selectivity.[104] Méndez-Lucio et al.[105] reported a generative model which links systems biology to molecular design. Their model was based on the conditional generative adversarial networks,[31] in which specific gene expression signatures were used as conditions. In this way, molecules with similar activity can be designed for the desired target without annotating the target of the compound.[105] Hong et al.[47] proposed conditional generation model CARAE based on ARAE,[79] in which they adopted a variational mutual information minimization framework to generate molecules with specific target properties. The original molecular properties were predicted by a predictor network, and the molecular properties can be separated from the latent vectors by minimizing the variational mutual information. In the decoding phase, the molecular structure is reconstructed according to the latent vector and separated target property information.[47] Kotsias et al.[103] constructed a conditional recursive neural network (cRNN) to generate molecules meeting the required conditions. In the RNN-based molecule generation process, the different molecular descriptors are entered into the model as conditions. Specifically, the authors built two cRNN models: the PhysChem-based model used LogP, TPSA, MW, HBA, HBD, and quantitative estimation of drug-likeness (QED) concatenated with soft labels predicted by the QSAR model as conditions, while the FingerPrint-based model used Morgan fingerprints as conditions. In this method, conditional seed can direct the focus of the RNN to a specific subset of the chemical domain, such as biologically active compounds related to a specific protein target.[103] The transformer-decoder model LigGPT[49] also can be trained conditionally to generate molecules with specific properties or desired scaffolds. Specifically, the property conditions and scaffold conditions are transformed into condition vectors and concatenated with the embeddings obtained from SMILES tokens. Therefore, the token predicted by a well-trained model can learn from both the previous tokens and the conditions. Due to the urgency of the COVID-19 pandemic, Chenthamarakshan et al.[106] proposed a generative model called controlled generation of molecules (CogMol) to design molecules targeting novel viral proteins with a set of desired attributes, by introducing a multiattribute controlled sampling scheme into the VAE model. They used CogMol to generate novel molecules for three SARS-CoV-2 target proteins, the main protease, the receptor-binding domain of the spike protein, and nonstructural protein 9 replicase, with constraints of target affinity and selectivity, drug-likeness, synthesis feasibility, and toxicity. The result showed that generated molecules were able to bind favorably to the relevant druggable pockets of the target structures and showed low predicted metabolite toxicity and high synthetic feasibility.[106]

The conditional generative models have also been applied to the generation of molecular graphs. Li et al.[51] proposed a conditional generative model based on MolRNN to generate graph molecules, which is suitable for multiobjective de novo drug design. In this study, they applied this model to generate compounds containing a given scaffold, based on synthetic accessibility and drug-likeness, as well as dual inhibitory activities against both JNK3 and GSK-3β.[51]

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

Table 3. De Novo Molecular Design Tools Based on Deep Generative Models

model	URL	developer	references
MolAICal	https://molaical.github.io/	Key Lab of Preclinical Study for New Drugs of Gansu Province, Lanzhou University	107
MORLD	http://morld.kaist.ac.kr	Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology	108
LiGANN	https://www.playmolecule.org/LiGANN/	Computational Science Laboratory, Universitat Pompeu Fabra	109

De Novo Molecular Design Tools Based on Generative Models. Many web-based applications and software that integrate generative models have been established, which are friendly to users that have less experience in coding or the knowledge of artificial intelligence. Table 3 summarizes a few publicly available tools.

MolAICal is a free and effective software developed by Bai et al., which can be used to design 3D ligands in the protein pocket. It contains sequence-based generative models and graph-based generative models, and both of them are trained by the Wasserstein generative adversarial network (WGAN).[107] MORLD is a free web server that can be used to automatically optimize lead compounds, and it can also generate small molecules from scratch targeting specific proteins when feeding the structure of protein without lead compound information.[108] LiGANN is a web-based application, and users only need to input the target protein PDB file for ligand generation. Different from the conventional generative models that learn from drug-like or target-specific data sets to generate new molecules, LiGANN is based on a new approach for structure-based drug design, where the neural network model was trained to map protein structures to ligand shape and then decode the shape to ligand in the form of SMILES.[109]

BENCHMARKS AND METRICS FOR A GENERATIVE MODEL

There are two main benchmarks for de novo molecular design, molecular sets (MOSES)[110] and GuacaMol,[111] which cover frequently used generative models and various metrics to evaluate the performance of generative models.

MOSES[110] is primarily concerned with the problem of evaluating the distribution of generated molecules. It contains five neural-network-based baseline models, namely, CharRNN,[16] VAE,[26] AAE,[29] JT-VAE,[57] and LatentGAN,[78] and three non-neural baselines, namely, the n-gram generative model, the hidden Markov model, and the combinatorial generator. Its data set is derived from ZINC.[112] In MOSES, "validity", "uniqueness", and "novelty" are the three most widely used metrics to evaluate the quality of molecules generated by various models. "Validity" describes the percentage of SMILES that can be recognized by RDKit in generated molecules, "uniqueness" represents the proportion of non-redundant molecules in valid molecules, and "novelty" is the fraction of generated molecules that are not in the training set. Other metrics used in MOSES are shown in Table 4.

GuacaMol[111] provides seven baseline models for generating molecules, including SMILES genetic algorithms,[113] graph genetic algorithms,[8] graph MCTS,[8] SMILES LSTM,[16] VAE,[41] AAE,[104] and ORGAN,[38] and the postprocessed ChEMBL[114] data set for training. The performances of these models are compared in two different aspects: one is that the model generates new molecules following the same chemical distribution of the training set, and the other is that generated molecules can meet specific properties. Correspondingly,

Table 4. A List of Performance Metrics for Molecular Generative Models

metrics	description
validity	The proportion of SMILES that can be recognized by RDKit in generated molecules.
uniqueness	The proportion of non-redundant molecules in valid molecules.
novelty	The proportion of generated molecules that are not in the training set.
filters	The proportion of valid molecules that can pass the custom medicinal chemistry filters and PAINS[116] filters.
fragment similarity	Compare the frequencies of BRICS fragments[117] in the generated set and the test set.
scaffold similarity	Compare the frequencies of Bemis–Murcko scaffolds[118] in the generated set and the test set.
similarity to a nearest neighbor	The average similarity between the generated molecule and the nearest neighbor molecule in the test set.
internal diversity	The similarity within the generated molecules.
Fréchet ChemNet distance (FCD)[115]	Fréchet distance between the distribution of the training set and the generated molecules.
KL divergence	Compare the probability distributions of the physicochemical descriptors from the molecules in the training set and the generated molecules.

metrics for these two aspects are also considered. For the distribution-learning benchmarks, three general metrics, "validity", "uniqueness", and "novelty", are assessed, and "FCD"[115] is also used in GuacaMol (Table 4). In addition, "KL divergence"[27] is used to compare the probability distributions of the physicochemical descriptors of the training molecules and the generated molecules (Table 4). For the goal-directed benchmarks, there are several different categories of optimization goals, such as rediscovering the target molecule, generating molecules similar to the target molecule, generating isomers corresponding to the target molecular formula, and so on.[111]

GraphINVENT[119] is a new benchmark for molecular graph generation, which integrates six different advanced GNN into a unified graph generation model framework, including message neural network, gated-graph neural network (GGNN),[89] set2vec (S2V),[120] GGNN with attention,[121] S2V with attention,[121] and edge memory network.[122] The metrics in this benchmark follow the MOSES, and the comparison results showed that GGNN performed best among all six GNN-based models.[119]

In addition to the metrics that have been introduced into the above benchmarks, Zhang et al. proposed that, when training the generative model with a small subset of GDB-13, the fraction of structures, ring systems, and functional groups of sampled molecules appearing in GDB-13 could be used to measure the chemical space coverage of generated molecules.[123]

CONCLUSION AND PROSPECT

De novo drug design is a process with a long cycle and high investment. With the fast progress of artificial intelligence, more

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

and more related approaches have been proposed.[124,125] Generative models have attracted our attention and developed rapidly, and different architectures that have been successful in other fields such as image or text generation have been proposed to generate new lead compounds with expected biological and chemical properties. In this Perspective, we mainly summarize recently reported generative modeling techniques and demonstrate their applications in the field of de novo drug design.

Although there have been a lot of studies on generative models for generating molecules, the application of generative models in drug design is still in its infancy, and there are many challenges to be addressed for further development.

For the purpose of extending the existing compound library, there have been many virtual libraries containing valid and novel chemical structures, including the generic database (GDB)[126−128] by the Reymond laboratory, ZINClick,[129] REAL,[130] DrugspaceX,[131] and so on. These libraries are either generated by predefined rule-based transformations or from mathematical graphs irrespective of pre-existing building blocks. Different generation approaches may have different advantages, such as synthetically more accessible or structurally more diversified. There have been some examples of successful discovery of new active ligands from these compound libraries through virtual screening.[132−134] For the deep generative model, an apparent advantage is it can be trained to learn a joint probability distribution of molecular representation and associated property, which allows us to more effectively sample new molecules that meet certain properties. There have been some reported works that try to explore the chemical space to obtain molecules that meet certain physicochemical properties of molecules,[101−103] and this is an emerging direction that needs to be further explored. Recently, Polykovskiy et al. used the model ECAAE to generate selective inhibitors of JAK3 by specifying high activity against JAK3 and low activity against JAK2 as a condition, and a hit with IC50 activity of 6.73 μM against JAK3 but inactivity against JAK2 was obtained.[104]

In terms of molecular representation in the generative model, many efforts have been devoted to study molecular topological graphs, but their performances often lack comparability due to different data sets and metrics used. GraphINVENT[119] is a benchmark test specially designed for comparing molecular graph generative models, while it does not include many new models. We expect that, with the improvement of benchmarking methods, the comparison among different generative models will become more standardized and more objective. Furthermore, we also see that attempts are being made to add information about the three-dimensional chemical structures, aiming to describe the structure of the molecule more accurately, thus making the molecules generated by the models more reliable for further research.

The currently widely used performance metrics for generative models also need to be improved. For example, Renz et al. have shown that "validity" can be easily maximized by inserting a carbon atom into SMILES strings in the training set.[135] According to the definition of "novelty", a molecule is novel if its SMILES is different from those in the training set, and this calculated "novelty" is different from the understanding of chemists, and many reported generative models have shown fairly high numerical values of "novelty".[110] The frequently used "druggability" and "synthesizability" metrics also have their own problems.[110] Therefore, although different evaluation and comparison metrics for generating models have been provided, the role and importance of these metrics for different studies are still unclear. How to evaluate the quality of a model and the generated molecules remains an unsolved issue, which requires a concerted effort to better refine baseline evaluation methods and to evaluate the ability of published generative models.

Another apparent shortcoming of the existing research is the lack of experimental verification. Although there have been many reports of using generative models to generate new compounds, there are less examples that generated compounds have been synthesized and experimentally evaluated. Zhavoronkov et al.[56] used the molecule GENTRL to discover effective inhibitors of DDR1 within 21 days. They designed, synthesized, and experimentally verified the molecule targeting DDR1 kinase in less than 2 months and finally obtained a drug candidate with good pharmacokinetic properties in experimental animals. This successful case illustrates the feasibility of the generative model for rapid drug design, but we also need to be cautious because the generated molecules are still in the early stage of drug development and may require further evaluation of their efficacy and safety in humans. Moreover, the similarity between the most active compound reported in this work and the known kinase inhibitor Ponatinib raises the question of whether similar active molecules can be generated using conventional molecular optimization strategies, such as fragment substitution, bioisosterism, rearrangement of heterocyclic ring systems, and so on.[136] This study also reminds us that the novelty of the generated molecules needs to be critically evaluated when applying the generative model to drug design.

Overall, we are just beginning to use generative models to design molecules, there are many aspects of such models that need to be further improved, and more computational and experimental validations and benchmarking tests are needed. Nevertheless, we believe that it will become an important pillar in the field of de novo drug design in the near future, assisting medicinal chemists to generate new ideas and accelerate the cycle of drug discovery.

AUTHOR INFORMATION

Corresponding Authors

Hualiang Jiang − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China*; Phone: +86-21-50806600-1303; Email: hljiang@simm.ac.cn

Nan Qiao − *Laboratory of Health Intelligence, Huawei Technologies Co., Ltd, Shenzhen 518100, China*; Phone: +86-15810851722; Email: qiaonan3@huawei.com

Mingyue Zheng − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China*; orcid.org/0000-0002-3323-3092; Phone: +86-21-50806600-1308; Email: myzheng@simm.ac.cn

Authors

Xiaochu Tong − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Xiaohong Liu − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia*

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China

Xiaoqin Tan − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Xutong Li − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Jiaxin Jiang − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Zhaoping Xiong − *Laboratory of Health Intelligence, Huawei Technologies Co., Ltd, Shenzhen 518100, China*

Tingyang Xu − *Tencent AI Lab, Shenzhen 518057, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.1c00927

Author Contributions

H.J., N.Q., and M.Z. directed the project. X.T. wrote the manuscript with the assistance of X.L., X.T., X.L., J.J., Z.X., and T.X. All authors approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Biographies

Xiaochu Tong is currently a postgraduate student in the Shanghai Institute of Materia Medica (SIMM), Chinese Academy of Sciences. Her research interests are focused on the application of generative models in drug design.

Xiaohong Liu is currently a Ph.D. student at SIMM, Chinese Academy of Sciences. His recent research is on artificial intelligence (AI)-based molecular filtering models for existing compound libraries or AI-designed molecules.

Xiaoqin Tan is currently a Ph.D. student at SIMM, Chinese Academy of Sciences. Her research is about molecular design and optimization based on generative models.

Xutong Li is currently a Ph.D. student at SIMM, Chinese Academy of Sciences. Her research interests are mainly focused on kinome-wide polypharmacology profiling of small molecules based on a multitask deep neural network.

Jiaxin Jiang was awarded a master's degree by University of Science and Technology of China. He is currently a staff member at the State Key Laboratory of Drug Research at SIMM. His main research interest is the development of artificial intelligence drug design methods, including but not limited to lead compound discovery and molecular generation based on machine learning methods.

Zhaoping Xiong graduated from ShanghaiTech University as a Ph.D. in 2021 and joined the Healthcare and Health Intelligence department of HUAWEI Cloud. His research interests are molecular representation learning, explainable AIs, and federated learning for drug discovery.

Tingyang Xu is a senior researcher at Machine Learning Center in Tencent AI Lab. His main research interests include social network analysis, graph neural networks, and graph generations, with particular focus on the deep graph learning models for molecular generation.

Hualiang Jiang was awarded a Ph.D. by SIMM, Chinese Academy of Sciences, in 1995. Dr Jiang mainly focuses on research about computational chemistry/biology and drug design. He has been engaged in the establishment of an innovative drug research platform by integrating target discovery and drug design methods and technologies.

Nan Qiao received his Ph.D. degree from Chinese Academy of Sciences in Bioinformatics and focuses on research about bioinformatics, genomics, clinical research, drug discovery, big data, machine learning, and artificial intelligence. He made significant contributions to cancer drug development during his stay at Novartis and won the Novartis Team Award and Novartis Select Award. In 2015, Nan joined Accenture China as the lead data scientist and set up Accenture China AI Lab, which focuses on building AI capabilities, AI assets, and AI industry solutions. In 2019, Nan joined HUAWEI Cloud as Chief Scientist in Healthcare and the head of Health Intelligence, leading AI products/services development for health industry.

Mingyue Zheng received his Ph.D. degree from SIMM under Professor Hualiang Jiang in 2006. He is currently a Professor in State Key Laboratory of Drug Research at SIMM. His main research interests are in artificial intelligence approaches for rational drug design and discovery, cheminformatics, and computational biology. He has been engaged in the machine-learning-based methodology development around the discovery and structural optimization of lead compounds, the assessment of drug ADME/T properties, as well as the application of the above methods in practical drug design and discovery processes.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81773634), Shanghai Municipal Science and Technology Major Project, National Science & Technology Major Project "Key New Drug Creation and Manufacturing Program" of China (Number: 2018ZX09711002-001-003), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12020372), and Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202002).

■ ABBREVIATIONS USED

AAE, adversarial autoencoder; AE, autoencoder; AI, artificial intelligence; ARDs, antibiotic resistance determinants; ARAE, adversarially regularized autoencoder; AT, adversarial threshold; ATNC, adversarial threshold neural computer; CB, cannabinoid; CDK4, cyclin-dependent kinase 4; CFG, context-free grammar; CNLL, conditional negative log-likelihood; CNN, convolutional neural network; CogMol, controlled generation of molecules; ConditionalGAN, conditional generative adversarial network; ConditionalVAE, conditional variational autoencoder; cRNN, conditional recursive neural network; dcGAN, deep convolutional generative adversarial network; DDR1, discoidin domain receptor 1; DNC, differentiable neural computer; DRD2, dopamine receptor type 2; druGAN, drug generative adversarial network; ECAAE, entangled conditional adversarial autoencoder; EDG, Euclidean distance geometry; FCD, Fréchet ChemNet distance; GAN, generative adversarial network; GCPN, graph convolutional policy network; GDB, generic database; GENTRL, generative tensorial reinforcement learning; GGNN, gated-graph neural network; GNNs, graph neural networks; GPCRs, G-protein coupled receptors; GPT, pre-training transformer; GRAPHDG, graph distance geometry; GRUs, gated recurrent units; GrammarVAE, grammar variational autoencoder; HBAs, hydrogen bond acceptors; HBDs, hydrogen bond donors; JAK3, Janus kinase 3; KL, Kullback−Leibler; LSTM, long short-term

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

memory; MACCS, molecular access system; MCTS, Monte Carlo tree search; MOSES, molecular sets; MW, molecular weight; ORGAN, objective-reinforced generative adversarial network; ORGANIC, objective-reinforced generative adversarial network for inverse-design chemistry; Pim1, Moloney murine leukemia virus kinase 1; PPAR, peroxisome proliferator-activated receptor; QED, quantitative estimation of drug-likeness; RANC, reinforced adversarial neural computer; RL, reinforcement learning; RNN, recurrent neural network; RXR, retinoid X receptors; S2V, set2vec; SA, synthetic accessibility; SD-VAE, syntax-direct variational autoencoder; SMILES, simplified molecular input line entry system; SSVAE, semi-supervised variational autoencoder; TPSA, total polar surface area; VAE, variational autoencoder; WGAN, Wasserstein generative adversarial network

REFERENCES

(1) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health. Econ.* **2016**, *47*, 20−33.

(2) Gaulton, A.; Kale, N.; van Westen, G. J. P.; Bellis, L. J.; Bento, A. P.; Davies, M.; Hersey, A.; Papadatos, G.; Forster, M.; Wege, P.; Overington, J. P. A Large-Scale Crop Protection Bioassay Data Set. *Sci. Data* **2015**, *2*, 150032.

(3) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(4) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.

(5) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(6) Kumar, A.; Voet, A.; Zhang, K. Y. Fragment Based Drug Design: From Experimental to Computational Approaches. *Curr. Med. Chem.* **2012**, *19*, 5128−5147.

(7) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079−1087.

(8) Jensen, J. H. A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chem. Sci.* **2019**, *10*, 3567−3572.

(9) Sheridan, R. P.; Kearsley, S. K. Using A Genetic Algorithm To Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.

(10) Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. 2015, arXiv:1511.06434. arXiv.org e-Print archive. https://arxiv.org/abs/1511.06434 (accessed Oct 22, 2020).

(11) Bowman, S.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; Bengio, S. Generating Sentences from A Continuous Space. 2015, arXiv:1511.06349. arXiv.org e-Print archive. https://arxiv.org/abs/1511.06434 (accessed Oct 22, 2020).

(12) Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. 2016, arXiv:1609.03499. arXiv.org e-Print archive. https://arxiv.org/abs/1609.03499 (accessed Dec 5, 2020).

(13) Engel, J.; Resnick, C.; Roberts, A.; Dieleman, S.; Eck, D.; Simonyan, K.; Norouzi, M. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. 2017, arXiv:1704.01279. arXiv.org e-Print archive. https://arxiv.org/abs/1704.01279 (accessed Dec 5, 2020).

(14) Schwalbe-Koda, D.; Gómez-Bombarelli, R. Generative Models for Automatic Chemical Design. 2019, arXiv:1907.01632. arXiv.org e-Print archive. https://arxiv.org/abs/1907.01632 (accessed Sept 10, 2020).

(15) Mikolov, T.; Karafiát, M.; Burget, L.; Cernocký, J.; Khudanpur, S. Recurrent Neural Network Based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*; 2010; Vol. 2, pp 1045−1048.

(16) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(17) Gupta, A.; Muller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111.

(18) Bjerrum, E. Molecular Generation with Recurrent Neural Networks. 2017, arXiv:1705.04612. arXiv.org e-Print archive. https://arxiv.org/abs/1705.04612 (accessed July 29, 2021).

(19) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554−2558.

(20) Jordan, M. I. Attractor Dynamics and Parallelism in A Connectionist Sequential Machine. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society*; 1986; pp 531−546.

(21) Elman, J. L. Finding Structure in Time. *Cogn. Sci.* **1990**, *14*, 179−211.

(22) Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. *A Field Guide to Dynamical Recurrent Neural Networks*; IEEE Press: 2001.

(23) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735−1780.

(24) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 2014, arXiv:1412.3555. arXiv.org e-Print archive. https://arxiv.org/abs/1412.3555 (accessed Sept 10, 2020).

(25) Bengio, Y. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1−55.

(26) Kingma, D.; Welling, M. Auto-Encoding Variational Bayes. 2013, arXiv:1312.6114. arXiv.org e-Print archive. https://arxiv.org/abs/1312.6114 (accessed Sept 10, 2020).

(27) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.

(28) Kingma, D. P.; Mohamed, S.; Jimenez Rezende, D.; Welling, M. Semi-Supervised Learning with Deep Generative Models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*; 2014; Vol. 2, pp 3581−3589.

(29) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I. Adversarial Autoencoders. 2015, arXiv:1511.05644. arXiv.org e-Print archive. https://arxiv.org/abs/1511.05644 (accessed Sept 10, 2020).

(30) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. 2014, arXiv:1406.2661. arXiv.org e-Print archive. https://arxiv.org/abs/1406.2661 (accessed Sept 10, 2020).

(31) Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. 2014, arXiv:1411.1784. arXiv.org e-Print archive. https://arxiv.org/abs/1411.1784 (accessed Nov 3, 2020).

(32) Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S. Transformers: State-of-The-Art Natural Language Processing. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* **2020**, 38−45.

(33) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017; pp 5998−6008.

(34) Kaelbling, L. P.; Littman, M. L.; Moore, A. W. Reinforcement Learning: A Survey. *J. Artif Intell Res.* **1996**, *4*, 237−285.

(35) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*: MIT Press: 2018.

(36) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, 48.

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027

(37) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.* **2018**, *4*, No. eaap7885.
(38) Guimaraes, G.; Sanchez, B.; Farias, P.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. 2017, arXiv:1705.10843. arXiv.org e-Print archive. https://arxiv.org/abs/1705.10843 (accessed Sept 10, 2020).
(39) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J. L.; Chen, H.; Engkvist, O. Exploring the GDB-13 Chemical Space Using Deep Generative Models. *J. Cheminf.* **2019**, *11*, 20.
(40) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science (Washington, DC, U. S.)* **2018**, *361*, 360−365.
(41) Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using A Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(42) Sanchez, B.; Outeiral, C.; Guimaraes, G.; Aspuru-Guzik, A. Optimizing Distributions over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). 2018, ChemRxiv.5309668.v3. ChemRxiv Preprint. DOI: 10.26434/chemrxiv.5309668.v3 (accessed Sept 10, 2020).
(43) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 1175−1183.
(44) Kusner, M.; Paige, B.; Hernández-Lobato, J. Grammar Variational Autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*; 2017; Vol. *70*, pp 1945−1954.
(45) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. In *Proceedings of the 6th International Conference on Learning Representations*; 2018.
(46) Bjerrum, E. J.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8*, 131.
(47) Hong, S. H.; Ryu, S.; Lim, J.; Kim, W. Y. Molecular Generative Model Based on an Adversarially Regularized Autoencoder. *J. Chem. Inf. Model.* **2020**, *60*, 29−36.
(48) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2017**, *8*, 10883−10890.
(49) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. LigGPT: Molecular Generation Using A Transformer-Decoder Model. 2021, ChemRxiv.14561901.v1. ChemRxiv Preprint. DOI: 10.26434/chemrxiv.14561901.v1 (accessed Sept 10, 2020).
(50) Grechishnikova, D. Transformer Neural Network for Protein-Specific De Novo Drug Generation as A Machine Translation Problem. *Sci. Rep.* **2021**, *11*, 321.
(51) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective De Novo Drug Design with Conditional Graph Generative Model. *J. Cheminf.* **2018**, *10*, 33.
(52) Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained Graph Variational Autoencoders for Molecule Design. In *Proceedings of the 32nd International Conference on Neural Information Processing*; 2018; pp 7806−7815.
(53) Cao, N. D.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. 2018, arXiv:1805.11973. arXiv.org e-Print archive. https://arxiv.org/abs/1805.11973 (accessed Sept 10, 2020).
(54) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. 2018, arXiv:1806.02473. arXiv.org e-Print archive. https://arxiv.org/abs/1806.02473 (accessed Sept 10, 2020).
(55) Samanta, B.; De, A.; Jana, G.; Chattaraj, P.; Ganguly, N.; Rodriguez, M. NeVAE: A Deep Generative Model for Molecular Graphs. *In Proceedings of the 33rd AAAI Conference on Artificial Intelligence* **2019**, *33*, 1110−1117.
(56) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; 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*, 1038−1040.
(57) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*; 2018; Vol. *80*, pp 2323−2332.
(58) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60*, 1983−1995.
(59) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular Geometry Prediction Using A Deep Generative Graph Neural Network. *Sci. Rep.* **2019**, *9*, 20381.
(60) Simm, G. N. C.; Hernández-Lobato, J. M. A Generative Model for Molecular Distance Geometry. 2019, arXiv:1909.11459. arXiv.org e-Print archive. https://arxiv.org/abs/1909.11459 (accessed July 4, 2021).
(61) Simm, G. N. C.; Pinsler, R.; Hernández-Lobato, J. M. Reinforcement Learning for Molecular Design Guided by Quantum Mechanics. 2020, arXiv:2002.07717. arXiv.org e-Print archive. https://arxiv.org/abs/2002.07717 (accessed July 4, 2021).
(62) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(63) Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative Molecular Design in Low Data Regimes. *Nature Machine Intelligence* **2020**, *2*, 171−180.
(64) Yang, Y.; Zhang, R.; Li, Z.; Mei, L.; Wan, S.; Ding, H.; Chen, Z.; Xing, J.; Feng, H.; Han, J.; Jiang, H.; Zheng, M.; Luo, C.; Zhou, B. Discovery of Highly Potent, Selective, and Orally Efficacious p300/CBP Histone Acetyltransferases Inhibitors. *J. Med. Chem.* **2020**, *63*, 1337−1360.
(65) Li, X.; Xu, Y.; Yao, H.; Lin, K. Chemical Space Exploration Based on Recurrent Neural Networks: Applications in Discovering Kinase Inhibitors. *J. Cheminf.* **2020**, *12*, 42.
(66) Tan, X.; Jiang, X.; He, Y.; Zhong, F.; Li, X.; Xiong, Z.; Li, Z.; Liu, X.; Cui, C.; Zhao, Q.; Xie, Y.; Yang, F.; Wu, C.; Shen, J.; Zheng, M.; Wang, Z.; Jiang, H. Automated Design and Optimization of Multitarget Schizophrenia Drug Candidates by Deep Learning. *Eur. J. Med. Chem.* **2020**, *204*, 112572.
(67) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37*, 1700153.
(68) Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning Artificial Intelligence on the De Novo Design of Natural-Product-Inspired Retinoid X Receptor Modulators. *Commun. Chem.* **2018**, *1*, 68.
(69) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the Macromolecular Targets of De Novo-Designed Chemical Entities through Self-Organizing Map Consensus. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 4067.
(70) Yoshimori, A.; Asawa, Y.; Kawasaki, E.; Tasaka, T.; Matsuda, S.; Sekikawa, T.; Tanabe, S.; Neya, M.; Natsugari, H.; Kanai, C. Design and Synthesis of DDR1 Inhibitors with A Desired Pharmacophore Using Deep Generative Models. *ChemMedChem* **2021**, *16*, 955−958.
(71) Hopcroft, J. E.; Motwani, R.; Ullman, J. D. *Introduction to Automata Theory, Languages, and Computation*, 3rd ed.; Pearson: New York, 2006.
(72) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.
(73) Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*; 2017; Vol. *31*, pp 2852−2858.
(74) Panaretos, V. M.; Zemel, Y. Statistical Aspects of Wasserstein Distances. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 405−431.

(75) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for De Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194−1204.
(76) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular De Novo Design. *Mol. Pharmaceutics* **2018**, *15*, 4386−4397.
(77) Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Colmenarejo, S. G.; Grefenstette, E.; Ramalho, T.; Agapiou, J.; Badia, A. P.; Hermann, K. M.; Zwols, Y.; Ostrovski, G.; Cain, A.; King, H.; Summerfield, C.; Blunsom, P.; Kavukcuoglu, K.; Hassabis, D. Hybrid Computing Using A Neural Network with Dynamic External Memory. *Nature* **2016**, *538*, 471−476.
(78) Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A De Novo Molecular Generation Method Using Latent Vector Based Generative Adversarial Network. *J. Cheminf.* **2019**, *11*, 74.
(79) Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; LeCun, Y. Adversarially Regularized Autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*; 2018; Vol. *80*, pp 5902−5911.
(80) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training; 2018, Preprint. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.
(81) Mandhana, V.; Taware, R. De Novo Drug Design Using Self Attention Mechanism. *In Proceedings of the 35th Annual ACM Symposium on Applied Computing* **2020**, 8−12.
(82) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - The Worldwide Chemical Structure Identifier Standard. *J. Cheminf.* **2013**, *5*, 7.
(83) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(84) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for De Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14*, 3098−3104.
(85) Bian, Y.; Wang, J.; Jun, J. J.; Xie, X.-Q. Deep Convolutional Generative Adversarial Network (dcGAN) Models for Screening and Design of Small Molecules Targeting Cannabinoid Receptors. *Mol. Pharmaceutics* **2019**, *16*, 4451−4460.
(86) Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *In Proceedings of the 27th International Conference on Artificial Neural Networks* **2018**, *11139*, 412−422.
(87) Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; Battaglia, P. Learning Deep Generative Models of Graphs. 2018, arXiv:1803.03324. arXiv.org e-Print archive. https://arxiv.org/abs/1803.03324 (accessed Sept 10, 2020).
(88) Johnson, D. D. Learning Graphical State Transitions. In *Proceedings of the 5th International Conference on Learning Representations*; 2017.
(89) Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated Graph Sequence Neural Networks. 2015, arXiv:1511.05493. arXiv.org e-Print archive. https://arxiv.org/abs/1511.05493 (accessed Nov 3, 2020).
(90) Samanta, B.; De, A.; Ganguly, N.; Gomez-Rodriguez, M. Designing Random Graph Models Using Variational Autoencoders with Applications to Chemical Design. 2018, arXiv:1802.05283. arXiv.org e-Print archive. https://arxiv.org/abs/1802.05283v1 (accessed Nov 3, 2020).
(91) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859−877.
(92) Kolda, T. G.; Bader, B. W. Tensor Decompositions and Applications. *SIAM Rev.* **2009**, *51*, 455−500.
(93) Ritter, H.; Kohonen, T. Self-Organizing Semantic Maps. *Biol. Cybern.* **1989**, *61*, 241−254.
(94) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*; 2017; Vol. *70*, pp 1263−1272.
(95) Havel, T. F. Distance Geometry: Theory, Algorithms, and Chemical Applications. *Encyclopedia of Computational Chemistry* **1998**, *120*, 723−742.
(96) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.
(97) Li, Y.; Pei, J.; Lai, L. Learning to Design Drug-Like Molecules in Three-Dimensional Space Using Deep Generative Models. 2021, arXiv:2104.08474. arXiv.org e-Print archive. https://arxiv.org/abs/2104.08474 (accessed May 10, 2021).
(98) Stewart, J. J. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(99) Husch, T.; Vaucher, A. C.; Reiher, M. Semiempirical Molecular Orbital Models Based on the Neglect of Diatomic Differential Overlap Approximation. *Int. J. Quantum Chem.* **2018**, *118*, No. e25799.
(100) Francesco, B.; Tamara, H.; Alain, C. V.; Markus, R. qcscine/sparrow: release 1.0.0 (version 1.0.0), 2019. DOI: 10.5281/zenodo.3244106.
(101) Lim, J.; Ryu, S.; Kim, J.; Kim, W. Molecular Generative Model Based on Conditional Variational Autoencoder for De Novo Molecular Design. *J. Cheminf.* **2018**, *10*, 31.
(102) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59*, 43−52.
(103) Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct Steering of De Novo Molecular Generation with Descriptor Conditional Recurrent Neural Networks. *Nature Machine Intelligence* **2020**, *2*, 254−265.
(104) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for De Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398−4405.
(105) Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J. De Novo Generation of Hit-Like Molecules from Gene Expression Signatures Using Artificial Intelligence. *Nat. Commun.* **2020**, *11*, 10.
(106) Chenthamarakshan, V.; Das, P.; Hoffman, S. C.; Strobelt, H.; Padhi, I.; Lim, K. W.; Hoover, B.; Manica, M.; Born, J.; Laino, T.; Mojsilovic, A. CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models. 2020, arXiv:2004.01215. arXiv.org e-Print archive. https://arxiv.org/abs/2004.01215 (accessed May 10, 2021).
(107) Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: A Soft Tool for 3D Drug Design of Protein Targets by Artificial Intelligence and Classical Algorithm. *Briefings Bioinf.* **2021**, *22*, No. bbaa161.
(108) Jeon, W.; Kim, D. Autonomous Molecule Generation Using Reinforcement Learning and Docking to Develop Potential Novel Inhibitors. *Sci. Rep.* **2020**, *10*, 22104.
(109) Skalic, M.; Sabbadin, D.; Sattarov, B.; Sciabola, S.; De Fabritiis, G. From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design. *Mol. Pharmaceutics* **2019**, *16*, 4282−4291.
(110) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644.
(111) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for De Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096−1108.
(112) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(113) Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-Based De Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **2018**, *47*, 1431−1434.
(114) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.
(115) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Frechet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 1736−1741.
(116) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(117) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* **2008**, *3*, 1503−1507.
(118) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(119) Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Jannik Bjerrum, E. Graph Networks for Molecular Design. *Mach Learn Sci. Technol.* **2021**, *2*, 025023.
(120) Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to Sequence for Sets. 2015, arXiv:1511.06391. arXiv.org e-Print archive. https://arxiv.org/abs/1511.06391 (accessed July 29, 2021).
(121) Lindelöf, E. *Deep Learning for Drug Discovery, Property Prediction with Neural Networks on Raw Molecular Graphs*; Chalmers University of Technology: 2019.
(122) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(123) Zhang, J.; Mercado, R.; Engkvist, O.; Chen, H. Comparative Study of Deep Generative Models on Chemical Space Coverage. *J. Chem. Inf. Model.* **2021**, *61*, 2572−2581.
(124) Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R. K. Artificial Intelligence in Drug Discovery and Development. *Drug Discovery Today* **2021**, *26*, 80−93.
(125) Burki, T. A New Paradigm for Drug Development. *Lancet Digit Health* **2020**, *2*, e226−e227.
(126) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.
(127) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(128) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(129) Massarotti, A.; Brunco, A.; Sorba, G.; Tron, G. C. ZINClick: A Database of 16 Million Novel, Patentable, and Readily Synthesizable 1,4-Disubstituted Triazoles. *J. Chem. Inf. Model.* **2014**, *54*, 396−406.
(130) REAL Database. https://enamine.net/compound-collections/real-compounds/real-database (accessed July 29, 2021).
(131) Yang, T.; Li, Z.; Chen, Y.; Feng, D.; Wang, G.; Fu, Z.; Ding, X.; Tan, X.; Zhao, J.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. DrugSpaceX: A Large Screenable and Synthetically Tractable Database Extending Drug Space. *Nucleic Acids Res.* **2021**, *49*, D1170−D1178.
(132) Bréthous, L.; Garcia-Delgado, N.; Schwartz, J.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. Synthesis and Nicotinic Receptor Activity of Chemical Space Analogues of N-(3R)-1-Azabicyclo[2.2.2]oct-3-yl-4-Chlorobenzamide (PNU-282,987) and 1,4-Diazabicyclo[3.2.2]-Nonane-4-Carboxylic Acid 4-Bromophenyl Ester (SSR180711). *J. Med. Chem.* **2012**, *55*, 4605−4618.
(133) Luethi, E.; Nguyen, K. T.; Bürzle, M.; Blum, L. C.; Suzuki, Y.; Hediger, M.; Reymond, J.-L. Identification of Selective Norbornane-Type Aspartate Analogue Inhibitors of the Glutamate Transporter 1 (GLT-1) from the Chemical Universe Generated Database (GDB). *J. Med. Chem.* **2010**, *53*, 7236−7250.
(134) Nguyen, K. T.; Luethi, E.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. 3-(Aminomethyl)piperazine-2,5-Dione as A Novel NMDA Glycine Site Inhibitor from the Chemical Universe Database GDB. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3832−3835.
(135) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On Failure Modes in Molecule Generation and Optimization. *Drug Discovery Today: Technol.* **2019**, *32*−*33*, 55−63.
(136) Walters, W. P.; Murcko, M. Assessing the Impact of Generative AI on Medicinal Chemistry. *Nat. Biotechnol.* **2020**, *38*, 143−145.

https://doi.org/10.1021/acs.jmedchem.1c00927
J. Med. Chem. 2021, 64, 14011−14027