Generative Model for Proposing Drug Candidates Satisfying Anticancer Properties Using a Conditional Variational Autoencoder

Sunghoon Joo, Min Soo Kim,* Jaeho Yang, and Jeahyun Park

ABSTRACT: Deep learning-based molecular generative models have successfully identified drug candidates with desired properties against biological targets of interest. However, syntactically invalid molecules generated from a deep learning-generated model hinder the model from being applied to drug discovery. Herein, we propose a conditional variational autoencoder (CVAE) as a generative model to propose drug candidates with the desired property outside a data set range. We train the CVAE using molecular fingerprints and corresponding G150 (inhibition of growth by 50%) results for breast cancer cell lines instead of training with various physical properties for each molecule together. We confirm that the generated fingerprints, not included in the training data set, represent the desired property using the CVAE model. In addition, our method can be used as a query expansion method for searching databases because fingerprints generated using our method can be regarded as expanded queries.

INTRODUCTION

Computer-aided drug design (CADD), an initial step in drug discovery, provides drug candidates that have the desired activity against biological targets of interest by the computational optimization of compound structures. If a compound modulates the target such that the disease is altered by a CADD, then this so-called hit will be refined to improve its safety and effectiveness and eventually become a drug candidate. Only one among more than 10,000 hits tested in early drug discovery may eventually result in a drug reaching the market. Discovering and introducing a new drug into the market typically requires at least 10 years of research and clinical development efforts as well as an expenditure of approximately 2.6 billion U.S. dollars. Therefore, establishing the appropriate pool of candidates is the most important step in drug discovery.

To reduce a pool of hit candidates, computational approaches have been used to obtain the ideal molecular structure from the desired functionality. For a direct mapping between the desired property and drug candidates, the inverse quantitative structure–activity relationship (QSAR) has been applied. Inverse-QSAR aims to obtain an explicit inverse mapping \( P_0(x) \rightarrow x \) from a set of functions \( P_k(x) \) expressing the properties \( k \) to a molecular descriptor space \( x \). One of the methods to address this issue is the evolutionary algorithm, which includes the transformation and crossover of a molecular structure by handcrafted mutation rules repeatedly to obtain optimized functional compounds. However, inverse-QSAR is not well defined because most of the descriptors in space \( x \) are not mapped back to corresponding molecular structures because the space is not continuously populated.

Recently, the increase in generative models based on deep neural networks has resulted in new computational molecular design techniques that enable researchers who wish to seek candidate compounds to narrow down the molecules of interest in a wide chemical space. A deep generative model can be applied immediately to recognize the relationship between the hidden patterns of molecular structures and their properties by summarizing a data distribution. A molecular design involving a deep generative model comprises two processes: learning the hierarchical distribution of feature representations in a data-driven manner and creating new molecules from mapped latent representations (latent space). Once the molecule is mapped to the latent space, gradient-based optimization can be performed to select the vector that is closest to the desirable property in a continuous representation. For a point in the latent space, the model can directly produce corresponding molecules. For generative models of various structures, such as the recurrent neural network (RNN) and convolutional neural network (CNN), many attempts have been undertaken to learn the chemical database in the form of computer readable chemical structures, such as SMILES, molecular graphs, and fingerprints (FPs) with physical properties. Recently, generative models via
Autoencoders\textsuperscript{16−18} have been discovered as sufficiently powerful for learning any type of data distribution using unsupervised learning to generate new data points with some variations.\textsuperscript{19−21} Accordingly, pharmaceutical scientists have employed autoencoder variants, such as VAE,\textsuperscript{19} generative adversarial network (GAN),\textsuperscript{20} and adversarial autoencoders,\textsuperscript{21} to randomly generate new molecules sampled from a continuous latent space of the molecules. Gómez-Bombarelli et al. demonstrated that a molecular space can be mapped to a continuous latent vector space and that chemical structures can be generated by sampling the latent vectors.\textsuperscript{22} Another advantage of the autoencoder-based molecular generative model is that it can be used in various molecular representations for training by changing the structure of the layer constituting the autoencoder. SMILES,\textsuperscript{22} MACCS FP,\textsuperscript{18,23} and molecular graphs were used as the input and output of autoencoders based on generative models. To generate novel molecules with the desired properties, the generative model has been extended to a conditional molecular design by an additional optimization procedure. Using the conditional variational autoencoder (CVAE), the molecular generation process can be conditioned on class-embedding vector $y$, which corresponds to the target properties.

\section*{Methods}

\textbf{Model Architecture and CVAE Training.} We adapted the CVAE to model the training data distribution in a continuous latent space and generate new FPs with the desired properties from the latent space (Scheme 1A). A CVAE is a

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{scheme1.png}
\caption{Outline of FP Generation and DB Screening\textsuperscript{a}}
\end{figure}

\textsuperscript{a}(A) Architecture of the CVAE in this study. FPs were used as the encoder input (orange boxes), and one-hot vectors of $-\log\text{GI50}$ were used as the class-embedding vector $y$ (green box). (B) Schematic illustration of the training procedure of the CVAE model. For each epoch, loss function including KL divergence and reconstruction loss was calculated. (C) Procedure for generating FPs satisfying anticancer activity and screening of effective molecules within chemical DBs. For a random vector in the latent space, novel FPs were generated from the trained CVAE model. The process can be conditioned on class-embedding vector $y$, which corresponds to the target properties.

Despite the continuing developments of these techniques, the low rate of validity suggests fundamental issues for practical applications in drug discovery.

Herein, we propose a generative model to generate new MACCS FPs satisfying anticancer properties as seeds to search candidates for hit compounds. We trained a CVAE\textsuperscript{29} on a data set from NCI-60 drug screening\textsuperscript{30} to generate molecular FPs from the latent representation, which contains information regarding the structure–activity relation from bioscreen data. Building a deep generative model based on compact feature vectors such as MACCS FPs are advantageous for real environments that are limited to small data sets.\textsuperscript{31} We utilized normalized GI50 (growth inhibition by 50\%\textsuperscript{a}) growth inhibition results as one-hot conditional vectors to generate molecular FPs that exhibit anticancer activity. GI50 is a robust metric of the efficacy of drugs in combating cancer cells.\textsuperscript{32} Even though GI50 is suitable for representing the anticancer activity of each molecule by one number, no study has used GI50 as a class-embedding vector for conditional generative models used in the search for a novel molecule. The CVAE generated MACCS FPs conditioned on the desired ranges of GI50. From the generated FPs, we selected synthesizable molecules in public DBs such as PubChem,\textsuperscript{33} ZINC,\textsuperscript{34} and ChEMBL\textsuperscript{35} by similarity calculation. Meaningful compounds that exhibit anticancer activity in large public DBs can be filtered out rapidly without performing biological experiments.
supervised learning model that adds class-embedding vector \( y \) into the encoder and decoder networks of the VAE.\(^{29}\) Using the CVAE, the latent variable \( z \) and the class-embedding vector \( y \) with the decoder network can generate an \( x' \) similar to the desired property of \( y \). The CVAE uses three types of variables: input variable \((x, y)\) data set (molecule representation of \( x \) and class-embedding vector \( y \)) from that distribution. The output of the latent layer is a fully connected layer of the architecture of our CVAE model is shown in Scheme 1A. It comprises an encoder, a latent layer, and a decoder. The encoder network \((q_{\phi}(z|x, y))\) comprises a multilayer perceptron (MLP), which has two dense layers with 1000 neurons. The activation functions of the first and second dense layers of this encoder were the exponential linear unit (Elu) and tanh, respectively. The input to the encoder network was a 166-bit MACCS FP \((x)\) concatenated with a predefined layer vector \( y \) (one-hot representation of \(-\log GI50\)). The output of the encoder was interpreted as the mean and variance of a Gaussian distribution; subsequently, a latent layer \( z \) comprising a fully connected layer of the five dimensions was sampled from that distribution. The output of the latent layer concatenated with the class-embedding vector \( y \) then became an input of the decoder. The decoder network \((p_{\theta}(x|z, y))\) was also an MLP, which contained two dense layers with 1000 neurons and a final layer with 166 neurons. The activation functions of the first and second dense layers of this decoder were tanh and Elu, respectively. The sigmoid activation function was applied in the final layer. The output \( x' \) of the decoder represented 166-bit FPs. The goal of the CVAE is to maximize the marginal log-likelihood \( \log p_{\theta}(x|y) \) as follows:

\[
\log p_{\theta}(x|y) = E_{z \sim q_{\phi}(z|x, y)}[\log p_{\theta}(x|z, y)] = E_{z \sim q_{\phi}(z|x, y)}[\log p_{\theta}(x|z, y)] - E_{z \sim q_{\phi}(z|x, y)}[\log q_{\phi}(z|x, y)] \\
+ E_{z \sim q_{\phi}(z|x, y)}[\log q_{\phi}(z|x, y)] - D_{KL}(q_{\phi}(z|x, y)||p(z|x, y))
\]

Because the third term on eq 1 is always positive or zero, we need to train the network to maximize the log-likelihood by maximizing the following variational evidence lower bound:

\[
\log p_{\theta}(x|y) \geq E_{z \sim q_{\phi}(z|x, y)}[\log p_{\theta}(x|z, y)] - D_{KL}(q_{\phi}(z|x, y)||p(z|x, y)) = \mathcal{L}(x, y, \phi, \theta)
\]

**Generating FPs with Properties and Screening with Similarity Search.** To generate FPs, we fed the decoder with the class-embedding vector \( y \) and the sampled vector \( z \) from the Gaussian \( p(z) \sim \mathcal{N}(0, I) \) (Scheme 1C). The decoder generated 166-bit MACCS FPs; however, each FP does not necessarily represent a unique molecule. We indirectly verified the characteristics of the generated FPs by comparing them with the molecules listed in public chemical DBs (FDA-approved anticancer drugs for breast cancer, PubChem DB, ZINC12 clean drug-like DB, and ChEMBL DB). To measure the similarity between the generated FPs and compound DBs, we calculated the Tanimoto similarity coefficients,\(^{36}\) which is the most widely used method to compare two different molecules.\(^{37}\) To reduce the time for measuring Tanimoto similarity coefficients between generated FPs and over 90 million compounds in the DBs, we used the chemfp toolkit (version 1.5) because it can load FP files into the memory and search for molecules whose coefficient exceeds the threshold of the DBs.

**Data Set and Molecular Representation.** The deep generative model was trained on DB containing in vitro cytotoxicity results of different concentrations of approximately 100,000 compounds on 60 human cancer cell lines (NCI).\(^{38}\) Among the indicators of biological sensitivity, we used GI50 as the class-embedding vector associated with the input FPs in the training set. The GI50 value was the drug concentration resulting in a 50% reduction in the net proliferation rate during drug incubation compared with the same proliferation in control cell cultures.\(^{39}\) GI50 is a robust metric of the efficacy of drugs against cancer cells.\(^{40}\) Figure 1 shows a schematic illustration of converting from the DB of screening results for human tumor cell lines to the data set for training and testing a generative model. To express the molecular structures, we used the 166-bit MACCS FP, which is one of the binary vector representations.\(^{50}\) FPs, which are a binary vector representation, contain simplified chemical information of any chemical entity.\(^{41,42}\) Every bit in the vector means the existence (1) or absence (0) of some substructures of each molecule. The FP representation of chemical DBs can reduce storage space and increase calculation and search speed. Furthermore, it is easy and fast to convert compounds in chemical DBs, such as...
PubChem, and to obtain compounds in DBs similar to the generated FP. Moreover, the FP can be compressed sufficiently to be configured into an in-memory database to perform rapid similarity calculations with approximately 100 million PubChem FP databases. We converted the compounds in the NCI-60 data set into the MACCS FP format using the open-source cheminformatics toolkit RDkit.43 Next, growth inhibition values (−log GI50) for each compound were converted into one-hot vectors.

**RESULTS AND DISCUSSION**

**Conditional Generation of FPs from Trained CVAE Model.** First, we confirm that our conditional VAE model, which was trained on FPs with class-embedding vectors, can...
perform conditional molecule generation. As shown in Scheme 1B, randomly selected latent vectors and class-embedding vectors are necessitated to decode the vectors into FPs, which satisfy the desired anticancer properties. Because the activity of the generated FPs is difficult to identify directly, we indirectly confirmed the anticancer activity of the generated FPs by finding the normalized GI50 (−log GI50) of molecules similar to that of the generated FPs. The similarity coefficients between the generated FPs and compounds in the test set were calculated according to Tanimoto measures. The mean of the −log GI50 for molecules were similar.

Table 2. Number of Generated FPs Similar to FDA-Approved Drugs for Breast Cancer (60 Drugs)

| similarity | target property |
|------------|-----------------|
| 4 ≤ −log GI50 < 5 | 5 ≤ −log GI50 < 6 | 6 ≤ −log GI50 < 7 | 7 ≤ −log GI50 < 8 | 8 ≤ −log GI50 < 9 |
| >0.8 | 2203.4 ± 601.1 | 2899.8 ± 619.2 | 4293.2 ± 423.7 | 8936.6 ± 931.8 | 17379.8 ± 1824.0 |
| >0.85 | 533.2 ± 248.6 | 708.2 ± 233.7 | 1277.2 ± 223.4 | 3344.8 ± 398.3 | 730.1 ± 1120.9 |
| >0.9 | 98.4 ± 53.8 | 148.8 ± 61.5 | 343.4 ± 127.9 | 1178.0 ± 240.2 | 2911.0 ± 597.1 |
| >0.95 | 5.4 ± 4.5 | 10.6 ± 7.3 | 59.8 ± 32.8 | 222.4 ± 104.5 | 6160.5 ± 251.1 |

"The FPs were generated from 100,000 decoding attempts from the latent vectors and each conditional vector. The same procedure was repeated for the five different models.

Table 3. Number of Molecules in PubChem Exceeding Threshold Values (0.8, 0.85, 0.9, and 0.95) of the Similarity Coefficient (the Tanimoto Coefficient was Employed as a Threshold) for Given Three Generated FPs

| generated FP index | similarity index | assigned to generated FP |
|--------------------|------------------|-------------------------|
| (1)                | 0.8              | 952                     |
| (2)                | 0.85             | 20,875                 |
| (3)                | 0.9              | 6999                    |

"The total number of compounds in the PubChem DB was 96,470,329. The FPs generated from the model are shown in Figure 3.

Figure 3. Molecules in PubChem compound DB with high similarities to three generated FPs by CVAE for the class-embedding vector (target: 8 ≤ −log GI50 < 9).

Table 4. Number of Molecules in ZINC12 Drug-like DB that Exceeded Threshold Values (0.8, 0.85, 0.9, and 0.95) of the Similarity Coefficient (the Tanimoto Coefficient was Employed as a Threshold) for Three Generated FPs

| generated FP index | similarity coefficient | assigned to generated FP |
|--------------------|------------------------|-------------------------|
| (1)                | 0.8                    | 96                      |
| (2)                | 0.85                   | 7155                    |
| (3)                | 0.9                    | 238                     |

"The generated FPs is difficult to identify directly, we indirectly confirmed the anticancer activity of the generated FPs by finding the normalized GI50 (−log GI50) of molecules similar to that of the generated FPs. The similarity coefficients between the generated FPs and compounds in the test set were calculated according to Tanimoto measures. Table 1 shows that the mean of the −log GI50 for molecules were similar.
(Tanimoto similarity coefficients ≥0.9) to the molecules generated from the CVAE models conditioned on five different class-embedding vectors. We calculated the mean $-\log GI50$ of the generated FPs only when the FPs had similar molecules (similarity >0.9) in the test set. From the generated FPs, 6.2, 6.0, 5.8, 8.0, and 8.8% (from different class-embedding vectors: 4 ≤ $-\log GI50 < 5$, 5 ≤ $-\log GI50 < 6$, 6 ≤ $-\log GI50 < 7$, 7 ≤ $-\log GI50 < 8$, and 8 ≤ $-\log GI50 < 9$, respectively) of FPs have similar molecules in the test set (see Table S1). The conditional vectors corresponding to each range bin were assigned when the decoder generated FPs from the latent vectors. As the number of training epochs increased, the mean $-\log GI50$ values of similar molecules obtained in the test set approached the target values represented by the class-emb...
embedding vectors. In two cases (7 ≤ −log GI50 < 8 and 8 ≤ −log GI50 < 9), the mean −log GI50 values of similar molecules did not approach each other in the ranges. This may be indicative of insufficient training data sets corresponding to these ranges (see Table S2). However, the mean −log GI50 tended to increase as the conditional vectors implied higher −log GI50 values. Our CVAE model learned the underlying probability distribution of the features related with molecular structures conditioned on the class-embedding vector (anticancer activity). Therefore, for a class-embedding vector that corresponds to the target properties, the trained CVAE model reconstructed the novel FPs with the desired anticancer activity.

Additionally, we varied the range of −log GI50, i.e., the class-embedding vector, and studied its effect on the anticancer activity of the generated FPs. Using the five different ranges of −log GI50 (2, 1, 0.5, 0.2, and 0.1), we reproduced the class-embedding vectors of each molecule in the training set. After training the CVAE, 500 FPs were randomly generated for each class-embedding vector from the model. Subsequently, we extracted the mean −log GI50 value of molecules that were similar to the generated FPs in the test set and confirmed that the values matched the segment of the class-embedding vector. Figure 2 shows the mean −log GI50 of the molecules, which was highly similar to that of the FPs generated under various class-embedding vectors for conditional molecular generation. As the label increased, the means of −log GI50 of the screened molecules by the generated FP tended to increase in every case. Considering the means and standard deviations of each case, the best segment of the class-embedding vector was defined as 1 for the effective conditional generation of FPs.

Evaluation of Anticancer Activity of Generated FPs through Comparison with FDA-Approved Drugs for Breast Cancer. To evaluate the anticancer activity of conditionally generated FPs, we measured the similarity between the FPs and FDA-approved drugs for breast cancer.

In this evaluation, we generated 100,000 FPs of each of the conditions (−log GI50: 4−9) from the model, and the number of FPs with a high similarity (>0.9) to the approved drugs was counted. The same procedure was repeated for the five different models. Table 2 shows that the FPs generated with high growth inhibition values (8 ≤ −log GI50 < 9) as the conditional vector was more similar to the approved drugs than the FPs with lower target values for decoding. This result implies that conditional molecular generation using anticancer activity as class-embedding vectors can produce molecules with structures similar to the drugs that have anticancer effects. The two conclusions from our results are as follows: first, even though the growth inhibition value for the class-embedding vectors was obtained through an in vitro experiment, the generated FPs can map molecules that exhibited anticancer activity on the human body. Second, the level of growth inhibition rate in the molecular generation can yield a differential generation.

Comparison between Generated FPs and Public Chemical DBs. To evaluate the validity and anticancer activity of the generated FPs, we generated FPs satisfying a high −log GI50 (8 ≤ −log GI50 < 9) and calculated the Tanimoto similarity coefficient between the generated FPs and the PubChem DB. We selected three fingerprints that had low similarities with the training data set below the threshold. Through this experiment, we attempt to demonstrate that the CVAE-trained data distribution of the training set can generate FPs that are not included in the data set.

Based on a previous study, we set the similarity threshold to 0.8, which is the criteria for a low similarity. We generated 1000 FPs from three models, each satisfying 8 ≤ −log GI50 < 9. Among the 3000 FPs, 1586 (52.3%) were dissimilar (the similarity <0.8) from the training set. Finally, three representative FPs were used to look up similar compounds in the public databases. Table 3 shows that the number of molecules in PubChem that exceeded the threshold values for the similarity coefficient (the threshold values: 0.8, 0.85, 0.9, and 0.95). Figure 3 shows the three generated FPs from the CVAE model and the representative PubChem compounds that had the highest similarity to each generated FP. The molecules with a high similarity index to each generated FP had similar structures through visual inspection, even though a single FP can map multiple molecular structures. To confirm whether the use of class-embedding vectors to indicate high anticancer activity was well reflected in the generation results, we verified whether the closest molecules exhibited anticancer activity. Among the molecules similar to the generated FP (1), CID: 11609586 (Serdemelan) showed growth inhibitory activity against various human cancer cell lines. The use of CID: 59555617 for cancer inhibition has been patented. Finally, CID: 3457989, a compound similar to the generated FP (2), showed the potency of human tyrosyl-DNA phosphodiesterase 1 (Tdp1) inhibitory activity in a bioassay. Tdp1 inhibitors can be considered as candidates for potential anticancer drugs. Finally, CID: 320753 displayed remarkable in vivo anticancer activity for tumor model P388 Leukemia (intraperitoneal) in CD2F1 mice. Among the molecules that searched were similar (exceeded the 0.9 similarity index) to FP (1), FP (2), and FP (3) (17, 81, and 66 molecules, respectively), the number of molecules that were confirmed to exhibit activity against cancer cells or tumors according to in vivo or in vitro biologic test results as well as the number of molecules that were patented as anticancer agents were 11 (64.7%), 22 (27.2%), and 24 (36.4%), respectively. However, 0 (0.0%), 27 (33.3%), and 5 (7.6%) of the 164 molecules selected showed active responses against cancer cells or tumors from in vivo or in vitro results. Finally, 6 (35.3%), 32 (39.5%), and 37 (56.1%) of the 164 molecules selected could not confirm the bioactivity of the molecule because of the lack of biologic test results (see Tables S3–S5). To demonstrate that universal comparison is possible regardless of the type of chemical DBs by FP generation as a molecular descriptor, we compared the generated FPs to the ZINC12 drug-like and ChEMBL24 DBs. This procedure can be considered as a validation process to infer the structure of the generated FPs by comparing the FPs with different databases. The ZINC12 DB contains approximately 18 million drug-like compounds satisfying the criteria. Tables 4 and 5 show the number of molecules that exceeded the threshold levels for the similarity coefficient (the threshold values: 0.8, 0.85, 0.9, and 0.95) in the ZINC12 and ChEMBL24 DBs, respectively. The three FPs used in the previous experiment were used as a reference. Figure 4 shows the structure and characteristics of compounds with high Tanimoto similarity coefficients in the two chemical DBs. The molecules belonging to each group with a high degree of similarity to the generated FP (1) and FP (2) had similar structures, even though the molecules were from different public databases. When we attempted to obtain molecules similar to the generated FP (3) in the ZINC12 DB, we...
discovered that no molecules exceeded the 0.9 similarity index. This was expected as the ZINC12 DB is composed of molecules that are lower than 500 MW, whereas FP (3) is expected to represent molecules with large molecular weights according to similar molecules from PubChem and ChEMBL DBs.

The evaluation of the validity of the generated FPs from the model is the most challenging aspect in the computer-aided molecular design. We demonstrated that a simple approach to verify the validity of the generated FPs was by comparing them with a large chemical database. By comparing the generated FPs with the three public chemical DBs, we can select meaningful compounds that exhibit synthesizability and stability from various public chemical DBs using a simple FP generation with one biological characteristic. However, despite the selection of FPs with low similarity to the training set, the compounds in the public DB similar to the generated FP often exhibited a high similarity (similarity >0.8) to the training set. We investigated the similarity between the searched compounds, which were similar to the generated FPs (1–3) in the PubChem DB and the training set. The result shows that 4 (23.5%), 28 (34.5%), and 6 (9.1%) of the 164 molecules could not obtain a molecule higher with a similarity index exceeding 0.8 in the training set (see Tables S3–S5). Furthermore, we confirmed that the molecules similar with the generated FPs exhibited anticancer activity from the bioassay results or information from the literature. However, the proposed method is limited for decoding the generated FPs into the molecules directly. In future studies, we might adopt the RNN-based FP-to-molecule decoder from Kotsias et al.27 in our CVAE model. Although the proposed method has limitations, the CVAE for FPs shows potential as a simple and easy procedure for generating new molecules by leveraging the knowledge from smaller data sets, such as in-house biological screening results. We demonstrated that new molecules can be constructed using CVAE trained from approximately 30,000 small bioassay data sets. For a large DB containing more than 100 million molecules, a significant amount of time is required to compare the generated descriptors and all the molecules in the DB. This study was enabled by the use of FPs, a relatively simple molecular descriptor.

**CONCLUSIONS**

We demonstrated that a deep neural network-based generative model to propose MACCS FPs as drug candidates exhibited the desired property outside of the data set range; furthermore, we validated our method by comparing the generated FPs with molecules from public chemical DBs. Our results showed the following: (i) the model that learned the relationship between structural features of molecules through MACCS FPs and the GI50 as an indicator of anticancer activity of the molecules can serve as a molecular generator for producing FPs with the target –log GI50; (ii) from the public cheminformatics databases, new candidate compounds that have potential anticancer activity can be obtained. Because the verification of synthetic accessibility or reactivity is required in our screening method, this study may facilitate the next step in the molecular design of drugs.
Sequence Generation Models with KL-Control. In International Conference on Machine Learning: 2016.
(10) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. Sci. Adv. 2018, 4, No. eaap7885.
(11) 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.
(12) Ray, M. P.; Lake, R. E.; Moody, S. A.; Magdalova, V.; Sosolik, C. A Hyperthermal Energy Ion Beamline for Probing Hot Electron Chemistry at Surfaces. Rev. Sci. Instrum. 2008, 79, No. 076106.
(13) Oleczonek, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. Aust. J. Chem. 2017, 9, 48.
(14) Bjerrum, E. J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). arXiv 2017.
(15) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective de Novo Drug Design with Conditional Graph Generative Model. Aust. J. Chem. 2018, 10, 33.
(16) Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. In International Conference on Artificial Neural Networks; Springer: 2018, 412-422, DOI: 10.5121/csit.2012.2417.
(17) Blaschke, T.; Oleczonek, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. Mol. Inf. 2018, 37, 1700123.
(18) Kadurin, A.; Nikolaenko, 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.
(19) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. arXiv 2013.
(20) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Advances in neural information processing systems; Neural Information Processing systems: 2014, 67, 328.
(21) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial Autoencoders. arXiv 2015.
(22) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-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.
(23) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Comucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. Oncotarget 2017, 8, 10883−10890.
(24) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. J. Cheminf. 2018, 10, 31.
(25) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. J. Chem. Inf. Model. 2019, 59, 43−52.
(26) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskii, V.; Mamoshina, P.; Bozdganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. Mol. Pharmacaceutics 2018, 15, 4398−4405.
(27) Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct Steering of de Novo Molecular Generation Using Descriptor Conditional Recurrent Neural Networks (CRNNs). ChemRxiv 2019.
(28) Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; Kim, W. Y. Predicting Drug−Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. J. Chem. Inf. Model. 2019, 3981−3988.
(29) Kingma, D. P.; Mohamed, S.; Rezende, D. J.; Welling, M. Semi-Supervised Learning with Deep Generative Models. Advances in neural information processing systems; 2014, 4, 3581−3589.
(30) Shoemaker, R. H. The NCI60 Human Tumour Cell Line Anticancer Drug Screen. Nat. Rev. Cancer 2006, 6, 813−823.
(31) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design—a Review of the State of the Art. Mol. Syst. Des. Eng. 2019, 4, 828−849.
(32) Halfer, M.; Nipel, M.; Chung, M.; Sorger, P. K. Growth Inhibition Metrics Correct for Confounders in Measuring Sensitivity to Cancer Drugs. Nat. Methods 2016, 13, 521−527.
(33) Wang, Y.; Xiao, J.; Szeuk, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A Public Information System for Analyzing Bioactivities of Small Molecules. Nucleic Acids Res. 2009, 37, W623−W633.
(34) Irwin, J. J.; Shoichet, B. K. ZINC—a Free Database of Commercially Available Compounds for Virtual Screening. J. Chem. Inf. Model. 2005, 45, 177−182.
(35) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100−D1107.
(36) Jaccard index; https://en.wikipedia.org/wiki/Jaccard_index (accessed Jun 30, 2020).
(37) Bajusz, D.; Ráca, A.; Heberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? Aust. J. Chem. 2015, 7, 20.
(38) NCI-60 Growth Inhibition Data; https://wiki.nci.nih.gov/display/NCIDTPData/NCI-60+Growth+Inhibition+Data (accessed Jun 20, 2020).
(39) Boyd, M. R.; Paull, K. D. Some Practical Considerations and Applications of the National Cancer Institute in Vitro Anticancer Drug Discovery. Drug Dev. Res. 1995, 34, 91−109.
(40) Barnard, J. Chemical Structure Representation and Search Systems. Chemical Informatics Software & Consultancy Services; Sheffield, UK (November 18, 2003) 2003.
(41) Daylight Theory Manual; https://www.daylight.com/dayhtml/doc/theory/index.html (accessed Jun 20, 2030).
(42) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742−754.
(43) RDKit: Open-source cheminformatics Software; http://www.rdkit.org/ (accessed Jul 20, 2001).
(44) Titano, J. J.; Badgeley, M.; Schefflein, J.; Pain, M.; Su, A.; Cai, M.; Sinwinburne, N.; Zech, J.; Kim, J.; Bederson, J.; Mocco, J.; Drayer, B.; Lehar, J.; Cho, S.; Costa, A.; Oermann, E. K. Automated Deep-Neural-Network Surveillance of Cranial Images for Acute Neurologic Events. Nat. Med. 2018, 24, 1337−1341.
(45) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. Expert Opin. Drug Discovery 2015, 11, 137−148.
(46) Lacroix, J. F. A.; Meyer, C.; Ligny, Y. A. E.; Csoka, I. C. F.; Lacrampe, J. F. A.; Meyer, C.; Ligny, Y. A. E.; Csoka, I. C. F.; Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. Expert Opin. Drug Discovery 2015, 11, 137−148.