frontiers | Frontiers in Pharmacology

TYPE Opinion
PUBLISHED 13 October 2022
DOI 10.3389/fphar.2022.1046524

OPEN ACCESS

EDITED BY
Fenglin Liu,
Korea University of Technology and Education, South Korea

REVIEWED BY
Rajesh Kumar Pathak,
Chung-Ang University, South Korea
Nandan Kumar,
North East Institute of Science and Technology (CSIR), India

*CORRESPONDENCE
Qiaoyin Tan,
joytam@zjnu.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

RECEIVED 16 September 2022
ACCEPTED 03 October 2022
PUBLISHED 13 October 2022

CITATION
Kong W, Hu Y, Zhang J and Tan Q (2022), Application of SMILES-based molecular generative model in new drug design.
Front. Pharmacol. 13:1046524.
doi: 10.3389/fphar.2022.1046524

Application of SMILES-based molecular generative model in new drug design

Weiya Kong[1†], Yuejuan Hu[2†], Jiao Zhang[3] and Qiaoyin Tan[4*]

[1]School of Sports Medicine and Rehabilitation, Beijing Sport University, Beijing, China, [2]Nursing Department of Fenyang College of Shanxi Medical University, Fenyang, China, [3]Innovation and Entrepreneurship College of Hunan University of Finance and Economics, Changsha, China, [4]College of Teacher Education, Zhejiang Normal University, Jinhua, China

KEYWORDS

SMILES-based, molecular generative model, generative model, new drug design, drug design

1 Introduction

Drugs are playing an increasingly important role in the long struggle between man and disease. Drug discovery is the process of identifying potential new therapeutic entities and drug design is the process of finding new medications based on knowledge of biological targets involving the design of molecules (Zhou and Zhong, 2017). Drug discovery and design has been facing obstacles due to the large human, material and financial resources required. With the success of artificial intelligence in the fields of image processing, pattern recognition and natural language processing (Xie et al., 2022), deep generative model has attracted wide attention in the field of drug discovery, and it also shows a promising application prospect in the field of molecular design optimization. When a generative model is used to generate molecules, its essence is to learn the distribution of molecules in the training set, and then generate molecules similar to but different from those in the training set. Combined with evolutionary algorithm or reinforcement learning, the properties of the generated molecules can be further optimized (Tong et al., 2021; Tan et al., 2022a). The molecular representation in the generative model can be in many forms, including Simplified Molecular Input Line Entry System (SMILES), molecular graph, etc. Generative models can be roughly divided into five categories, including recurrent neural network, RNN, autoencoder, AE, generative aggressive network, GAN, Transformer and generative model combined with reinforcement learning, RL (Bhisetti and Fang, 2022) as shown in Figure 1A. Among them, the molecular generative model based on the text sequence (SMILES) is the most widely used. This paper simply introduces the basic principle and application of deep generative model based on the latest molecular design of the text sequence (SMILES), so that readers can understand deep generative model and use it better in drug molecular design.

2 Model

2.1 RNN-based model

Recurrent neural network (RNN) can accept sequence data as input features. It was used in natural language processing, but now it is used to generate new compound structures (Wang et al., 2021). When RNN model is used to generate molecules, the molecules can be expressed as the sequence SMILES. Because SMILES is a string sequence, it is very suitable to be processed by a neural network like RNN, as shown in Figure 1B. It is found that the RNN-based method can learn the low-dimensional distribution of molecular sequence grammar and chemical space with the SMILES representation of compound molecules as input. RNN has unique advantages for sequences with large differences in length distribution (Cheng et al., 2021). Segler et al. (Segler et al., 2018) and Olivecrona et al. (Olivecrona et al., 2017) use RNN network with long short-term memory (LSTM) and gated recurrent unit (GRU) to generate molecular structures. By using a large number of SMILES molecular structures as training sets, the RNN model they trained can automatically generate molecular structures with high drug-like properties and the efficacy is as high as over 90%, and the diversity of molecular structures obtained is basically the same as that of the training set. RNN first learns how a large number of SMILES texts represent molecules in a language-like way, and the fitted model can generate new SMILES strings, i.e., new molecules without bias, which are suitable for virtual screening and other applications.

2.2 VAE-based model

VAE consists of encoder and decoder. The research group uses convolutional neural network (CNN) as an encoder to map the input molecular structure into latent variables, while the decoder uses RNN to recover the hidden variables to the SMILES sequence corresponding to the original molecular structure, as shown in Figure 1C . Due to the randomness of VAE algorithm, different hidden variables in the hidden variable space can be sampled after training, which can then be decoded to obtain different molecular structures (Altae-Tran et al., 2017; Wu et al., 2018). In addition, VAE can encode high-dimensional data in low-dimensional space, and form a “feature space” in parallel, which we can also call “drug space”. To some extent, it represents the complete set of targeted drugs. Therefore, if we take another point in this area and decode it back to the high-dimensional chemical molecule, then this molecule is a potential targeted new drug. However, VAE has limitations: if all the medicine-ready molecules are used to construct the “drug space”, then the medicine re-sampled is only a new medicine-ready molecule with no targeting selectivity; If the “drug space” is constructed with drugs targeting a certain pocket, a large number of known drugs targeting this site are needed,

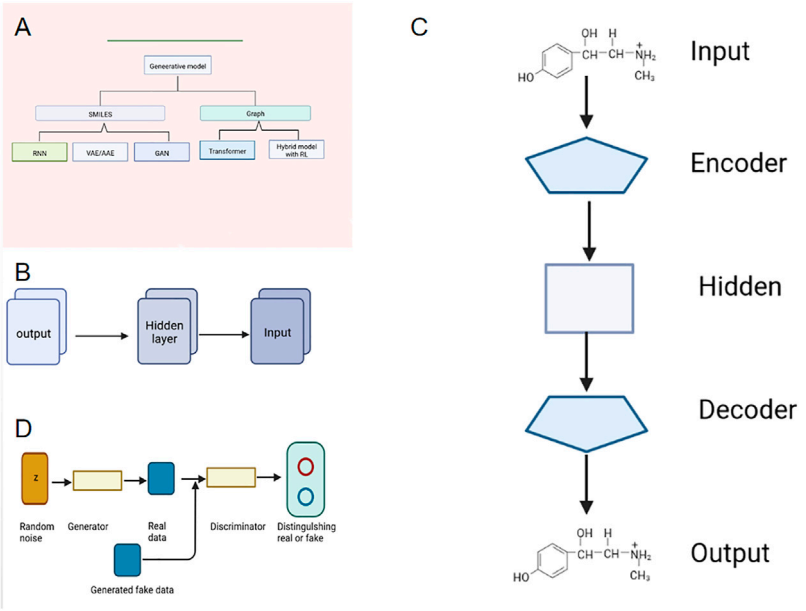

FIGURE 1
Different mode indication. **(A)** Generative Models for Molecular Design. **(B)** Structure of RNN. **(C)** Structure of VAE. **(D)** Structure of GAN.

which cannot be used in the development of first-class drugs (Lim et al., 2018; Ragoza et al., 2022).

2.3 GAN-based model

The concept of GAN was proposed by Goodfellow in 2014, which was inspired by the zero-sum game. It is a type of neural network used in unsupervised learning, which helps to solve tasks such as generating images by text, improving the resolution of images, matching drugs, and retrieving images with specific patterns (Tong et al., 2021; Wang et al., 2021; Tan et al., 2022a; Bhisetti and Fang, 2022). The model consists of two neural networks: the generator G used to fit the data distribution and the discriminator D used to judge whether the input is "true". In the training process, G outputs the hidden vector Z sampled from the prior distribution p(z) as a data space, while D distinguishes the real data from the output of the generated network as much as possible, thus forming a game process between the two networks to learn the production model of data distribution. Ideally, it will lead to a generative model that can be convincingly real (Kadurin et al., 2017), as shown in Figure 1D. In the molecular structure generation of GAN model, the generator generates random SMILES while the discriminator tries to distinguish these random SMILES from those of real molecules in the training set. In each round of training, the generator keeps learning, making the SMILES sequence generated by it closer and closer to the real molecule, until the discriminator cannot distinguish whether a SMILES sequence comes from the generator or the training set. Therefore, the generator network can be used to generate molecules (Polykovskiy et al., 2020).

3 Discussion

Some generative models have been successfully applied to generate new lead compounds with expected physical and chemical properties, but the application of generative models can be further explored in drug design.

Compared with the virtual compound library based on rules, the advantage of the generative model is that it can learn the joint probability distribution of molecular characterization and properties, which enables us to sample new molecules satisfying specific properties more effectively (Tong et al., 2021). Compared to the international chemical identifier (InCHI), which is also a one-dimensional linear representation, SMILES has a more rigorous syntax and uses a mapping algorithm from molecular graph to text (Elton et al., 2019). This makes SMILES easier to processing and more suitable for training machine learning models. The choice of SMILES as molecular input also does not suffer from the same limitations as fingerprints, i.e. the output is not directly converted into the true molecular structure and has difficulties in being used for *de novo* design. For generative models using 2D representations, i.e. molecular graph-based models, performance is often lacking in comparability due to different datasets and metrics; for generative models with 3D representations, they are limited to known molecular formulae only; whereas SMILES-based models are computationally lower cost, more easily scalable to larger molecules and/or larger datasets (Bilodeau et al., 2022), and can also benefit from improvements in algorithms related to natural language processing. Future research could translate specific target languages such as protein sequences into the SMILES language, i.e. the generation of molecules with specific characteristics could be considered as a translation. These methods may also be useful in bio drug design, such as stem cells (Tan et al., 2022b), growth factors (Tan et al., 2022c; Tan et al., 2022d), et al.

It is worth noting that despite the proliferation of SMILES-based models in recent years, it still has some limitations, such as the lack of explicit specification of molecular similarity, the possible inability to apply existing natural language processing models directly, and the need to additionally remove invalid SMILES. It is believed to be an important pillar in the field of new drug design in the near future, through continuous refinement to help pharmaceutical chemists expedite the process of drug discovery and design.

Author contributions

WK and YH contributed to conception and design of the study, and wrote the first draft of the manuscript. JZ contributed to the data collection and analysis. QT contributed to manuscript revision, read, and project management. All authors approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3 (4), 283–293. doi:10.1021/acscentsci.6b00367

Bhisetti, G., and Fang, C. (2022). Artificial intelligence-enabled de novo design of novel compounds that are synthesizable. *Methods Mol. Biol.* 2390, 409–419. doi:10.1007/978-1-0716-1787-8_17

Bilodeau, C., Jin, W., and Jaakkola, T. (2022). Generative models for molecular discovery: Recent advances and challenges[J]. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, e1608.

Cheng, Y., Gong, Y., Liu, Y., Song, B., and Zou, Q. (2021). Molecular design in drug discovery: A comprehensive review of deep generative models. *Brief. Bioinform.* 22 (6), bbab344. doi:10.1093/bib/bbab344

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular design—A review of the state of the art. *Mol. Syst. Des. Eng.* 4 (4), 828–849. doi:10.1039/c9me00039a

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). druGAN: An advanced generative adversarial autoencoder model for de Novo generation of new molecules with desired molecular properties *in silico*. *Mol. Pharm.* 14 (9), 3098–3104. doi:10.1021/acs.molpharmaceut.7b00346

Lim, J., Ryu, S., Kim, J. W., and Kim, W. Y. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminform.* 10 (1), 31. doi:10.1186/s13321-018-0286-7

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* 9 (1), 48. doi:10.1186/s13321-017-0235-x

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al. (2020). Molecular sets (moses): A benchmarking platform for molecular generation models. *Front. Pharmacol.* 11, 565644. doi:10.3389/fphar.2020.565644

Ragoza, M., Masuda, T., and Koes, D. R. (2022). Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* 13 (9), 2701–2713. doi:10.1039/d1sc05976a

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4 (1), 120–131. Epub 2017 Dec 28. doi:10.1021/acscentsci.7b00512

Tan, Q., Li, J., Liu, Y., Zhu, X., and Shao, W. (2022). Feasibility of growth factor Agent therapy in repairing motor injury. *Front. Pharmacol.* 13, 842775. doi:10.3389/fphar.2022.842775

Tan, Q., Li, J., Yin, Y., and Shao, W. (2022). The role of growth factors in the repair of motor injury. *Front. Pharmacol.* 13, 898152. doi:10.3389/fphar.2022.898152

Tan, Q., Wu, C., Li, L., Liang, Y., Bai, X., and Shao, W. (2022). Stem cells as a novel biomedicine for the repair of articular meniscus: Pharmacology and applications. *Front. Pharmacol.* 13, 897635. doi:10.3389/fphar.2022.897635

Tan, X., Li, C., Yang, R., Zhao, S., Li, F., Li, X., et al. (2022). Discovery of pyrazolo [3, 4-*d*]pyridazinone derivatives as selective DDR1 inhibitors via deep learning based design, synthesis, and biological evaluation. *J. Med. Chem.* 65 (1), 103–119. doi:10.1021/acs.jmedchem.1c01205

Tong, X., Liu, X., Tan, X., Li, X., Jiang, J., Xiong, Z., et al. (2021). Generative models for de novo drug design. *J. Med. Chem.* 64 (19), 14011–14027. doi:10.1021/acs.jmedchem.1c00927

Wang, Y., Wang, H., Yan, M., Hu, G., and Wang, X. (2021). Artificial intelligence design of biomolecular sequences [J]. *Synth. Biol.* 2 (1), 1–14.

Wu, Z. Q., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* 9 (2), 513–530. doi:10.1039/c7sc02664a

Xie, W., Wang, F., Li, Y., Lai, L., and Pei, J. (2022). Advances and challenges in de novo drug design using three-dimensional deep generative models. *J. Chem. Inf. Model.* 62 (10), 2269–2279. doi:10.1021/acs.jcim.2c00042

Zhou, S. F., and Zhong, W. Z. (2017). Drug design and discovery: Principles and applications. *Molecules* 22 (2), 279. doi:10.3390/molecules22020279