Commentary

For reprint orders, please contact: reprints@future-science.com

Future Medicinal Chemistry

Neural networks in drug discovery: current insights from medicinal chemists

Yinqiu Xu[1], Xuanyi Li[1], Hequan Yao[1] & Kejiang Lin*,[1]
[1] Department of Medicinal Chemistry, School of Pharmacy, China Pharmaceutical University, Nanjing, PR China
*Author for correspondence: link@cpu.edu.cn

> “Evaluations that use metrics aim to predict rather than to represent the real values of models, and the ability to use models for real tasks is what concerns medicinal chemists”

First draft submitted: 15 April 2019; Accepted for publication: 10 May 2019; Published online: 9 July 2019

Keywords: artificial intelligence • drug discovery • machine learning • neural networks

Neural networks in drug discovery

In recent years, neural networks (NNs) have become too effective to be mistaken as an equivalent of artificial intelligence (AI) [1]. Similar to the other machine learning (ML) algorithms for realizing artificial intelligence, NNs identify rules from samples. To some degree, NNs appear to be powerful mathematical formulas, which accurately describe the relationships between independent variables and dependent variables, and the training of NNs resembles the process of approximating the formulas. According to the universal approximation theorem, NNs with simple architecture can approximate many functions [2]. Though the main idea of NNs is easy to understand, the specific mechanisms are delicate and complicated (as previously described in [3,4]).

Furthermore, the relationships between the structure and activity in drug discovery can be described with formulas, which has resulted in the application of NNs to drug discovery. The dependent variables of drug discovery can be discrete (such as whether a compound is safe or not) or continuous (such as IC_{50} values), and the corresponding NNs can be trained as qualitative models or quantitative models. Due to the powerful ability of NNs to adapt, NNs always show better performance than other models (as previously reviewed in [1,4,5]). From the perspective of independent variables that describe molecules, NNs accept many ways of describing ligands [5], receptors and receptor–ligand interactions. In this way they cover the areas of ligand-based drug design, *de novo* drug design and receptor-based drug design.

The structures of molecules determine their properties, a concept which kept in mind by medicinal chemists and reflects the importance of molecular structures. Similarly, before preparing *in silico* models for drug discovery, how molecular structures are represented affects the quality of the models. Traditionally, molecular descriptors and fingerprints have been used to represent structures. However, they are preprocessed information rather than raw information of structures. Due to the preprocessing, molecular descriptors and fingerprints may not represent structures accurately and comprehensively. Though many kinds of NNs exist, such as genetic NNs, self-organizing maps and radial basis function NNs, they can hardly deal with the raw information of molecular structures. With the development of recurrent neural networks (RNNs), convolutional neural networks (CNNs) and molecular graph-based neural networks (MGNNs), raw information indicating molecular structures can be input into models based on NNs, which helps the machines analyze the rules directly and comprehensively.

Considering models based on NNs as tools for different tasks, the tools have been being developed from simple NNs to NNs with delicate architectures and diverse components. Artificial neural networks (ANNs) are the most basic components that combine different independent variables into more or fewer new variables, which can be further combined as subsequent variables or final outputs. RNNs are components that combine ANNs recursively so that their independent variables can be propagated dynamically, which make RNNs good at dealing with sequential inputs such as the simplified molecular input line entry specification (SMILES). Inside CNNs, different ANNs are used to combine local neighboring variables into different new variables. The idea of CNNs is shared by MGNNs, but MGNNs are implemented differently to adapt to non-Euclidean data. The components

newlands press

10.4155/fmc-2019-0118 © 2019 Newlands Press ISSN 1756-8919

mentioned above are being updated. Extra ANNs can be used to control, read and write the inner variables of NNs so that – to some degree – the updated NNs will have a memory. Meanwhile, the elaborate design of the inner elements of those NNs makes the NNs more adaptive. Take CNNs as an example. The dynamic k-max pooling provides CNNs with new potential to deal with sequential inputs, and the global average pooling further alleviates the overfitting and the effects of the input size. When the outputs of a single NN are set as new inputs rather than final results, the functions of models can be enlarged. The first NN can be set as an encoder, which converts its appropriate inputs into new representations and the new representations can be used by another NN for qualitative or quantitative tasks. More importantly, the new representations can be decoded by the latter NN, which can be used for generating new molecules. The simple process of encoding and decoding just regenerates the inputs, but introducing the adversarial mechanism and reparameterization makes the process different, which results in adversarial autoencoders and variational autoencoders. In addition to encoders and decoders, NNs can be set as generators and discriminators as well, which are two necessary parts of generative adversarial networks. After preparing the architecture of NNs, models that are more practical can be trained according to the learning methods. Transfer learning and one-shot learning are designed for tasks that have limited samples, and having enough samples is usually a problem in drug discovery. Reinforcement learning is a dynamic process during which models can be trained according to desired properties, and the process is exactly appropriate for generating SMILES sequences. The mentioned tools based on NNs accept raw information of molecules for traditional qualitative or quantitative tasks and connect in different ways, which provide colorful choices for drug discovery [5].

Reliability versus novelty

Appropriate architectures, optimized hyperparameters and accurate samples are three important factors for training good models. The first two factors are adjusted through careful validations, while the samples are usually prepared before training.

During the training process of NNs, the inner weights and biases of NNs are adjusted according to the inputted data. More data for a model means that the inner parameters of the model are more adaptive to various samples, so the model can be well generalized when dealing with new samples. Abundant samples make models reliable, which are needed by models to predict molecular properties such as water solubility and toxicity. However, a dilemma comes with an enlarged dataset, especially for medicinal chemists focusing on discovering new active compounds. In drug discovery, abundant existing samples for a target indicate that less space is available for further research and rigorous intellectual properties. Whether models based on NNs can discover desired molecules remains to be tested in practice.

Recently, a comparison between models based on NNs and other ML algorithms further emphasized the contradiction between reliability and novelty [6]. The models were found to provide worse predictions when dealing with newer samples, and models based on NNs did not provide much improvement when compared with several ML models. Well-designed architectures and optimized hyperparameters may not improve the performance of NNs much. The best solution is to supply stranger samples, which could result in the loss of novelty. Although there are models that are designed for tasks with limited samples, their performance remains uncertain. For instance, the performance of models based on one-shot learning was found to be unsatisfactory for virtual screening tasks [7]. Classifying and scoring docking results is also a solution to the contradiction. However, the NN-based models for these tasks significantly depend on the docking methods, and to some degree, the docking methods may limit the potential of these models.

From the perspective of medicinal chemists, NNs provide effective tools to choose compounds with better drug-likeness for clinical research, but more cases in practice are necessary to prove the ability of NNs in innovative tasks such as virtual screening and predicting potential targets.

More models *in silico* but limited applications in reality

Along with the fast growth of models, some details have been noticed to help improve models. For example, the effects of the hyperparameters [8], the quality of the dataset [9] and the diversity of SMILES sequences [10] have been mentioned. However, medicinal chemists may still get confused when choosing an appropriate model. Since sometimes complex models may not bring satisfactory results, comparing the metrics between different models will help alleviate the confusion, and benchmarks were proposed for traditional tasks [11] and *de novo* molecule generation [12].

Evaluations that use metrics aim to predict rather than to represent the real values of models, and the ability to use models for real tasks is what concerns medicinal chemists. Take the models for *de novo* molecule generation as an example. It is easy for these models to achieve satisfactory metrics, but the generated molecules may not be stable or chemically accessible. With respect to practical applications, solving a real problem in drug discovery is more important than evaluations *in silico* since drug discovery is never an easy task. Although most models were validated and tested theoretically, the good news is that there are now several successful cases that only depend on simple algorithms. Active compounds targeting nuclear receptors [13] and kinases [14,15] were discovered using NN-based models. Meanwhile, anticancer peptides were successfully designed using RNNs [16]. In these cases, models based on NNs prove their practical value.

NNs & automated drug discovery

It seems that NNs are now aiding drug discovery, but it is reasonable to foresee that models based on NNs will automatically perform drug discovery. Automatic devices for synthesis, analysis and biological tests represent a trend in the industry [17], though the devices just perform as they are designed. Combined with ML, those devices may determine their actions on their own so that an automatic ‘design-make-test’ cycle, which is a feedback loop, can be fulfilled in drug discovery [18]. It is worth noting that the loop has been successfully implemented using ML [19], and four reactions were successfully discovered. Among the diverse ML algorithms, NNs have proved their ability in many stages of drug discovery, such as the automatic design of new molecules and the design of synthetic routes [20]. Furthermore, the automatic cycle is also a dynamic process, which means that reinforcement learning may also help models based on NNs to learn rules and design new drugs. More importantly, NNs accept various kinds of data as inputs, which include both abstract information and intuitive representations such as words and images; further, compared with other ML algorithms, this characteristic enables NNs to analyze the results of their actions directly and conveniently. Based on those advantages of NNs, it is reasonable to infer that NNs may be better integrated into the automatic feedback loop in drug discovery.

Conclusion

Overall, it seems that medicinal chemists have not kept pace with the rapid updates in NN models. Although new ideas and models keep emerging for virtual drug discovery, it is hard to evaluate which one is the best for applications. More choices *in silico* do not represent the proper scope of the applications, and impressive practical cases are still limited. The main cause may be the uncertainty of NNs in dealing with difficult tasks with limited samples, which are usually the main focus of medicinal chemists. Cooperation between researchers in cheminformatics and medicinal chemists will help identify potential problems and create more effective models, which may result in bringing impressive results. Meanwhile, the cooperation will accelerate the integration of NNs and the ‘design-make-test’ cycle, so that automated drug discovery can be realized in the future.

Financial & competing interests disclosure

This work was supported by funds from the National Key R&D Program of China (2018YFC0311005), ’Double First-Class’ University Project, China (CPU2018GY15) and Key Laboratory of Spectrochemical Analysis & Instrumentation (Xiamen University), Ministry of Education, China (SCAI1802). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Writing assistance was utilized in the production of this manuscript. The authors thank the American Journal Experts (AJE), Durham, NC, USA, for their English language editing.

References

1. Carpenter KA, Cohen DS, Jarrell JT, Huang X. Deep learning and virtual drug screening. *Future Med. Chem.* 10(21), 2557–2567 (2018).
2. Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Net.* 4(2), 251–257 (1991).
3. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol. Inform.* 35(1), 3–14 (2016).
4. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov. Today* 22(11), 1680–1685 (2017).
5. Xu Y, Yao H, Lin K. An overview of neural networks for drug discovery and the inputs used. *Expert Opin. Drug Discov.* 13(12), 1091–1102 (2018).

6. Liu R, Wang H, Glover KP, Feasel MG, Wallqvist A. Dissecting machine-learning prediction of molecular activity: is an applicability domain needed for quantitative structure–activity relationship models based on deep neural networks? *J. Chem. Inf. Model.* 59(1), 117–126 (2019).

7. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3(4), 283–293 (2017).

8. Zhou Y, Cahya S, Combs SA *et al.* Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. *J. Chem. Inf. Model.* 59(3), 1005–1016 (2019).

9. Lieyang C, Anthony C, Steven R *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *ChemRxiv* doi:10.26434/chemrxiv.7886165.v1 (2019) (Epub ahead of print).

10. Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. *ArXiv E-prints* 1703.07076 (2017).

11. Wu Z, Ramsundar B, Feinberg EN *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9(2), 513–530 (2018).

12. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for *de novo* molecular design. *J. Chem. Inf. Model.* 59(3), 1096–1108 (2019).

13. Merk D, Friedrich L, Grisoni F, Schneider G. *De novo* design of bioactive small molecules by artificial intelligence. *Mol .Inform.* 37(1–2), 1700153 (2018).

14. Xu Y, Chen P, Lin X, Yao H, Lin K. Discovery of CDK4 inhibitors by convolutional neural networks. *Future Med. Chem.* 11(3), 165–177 (2018).

15. Polykovskiy D, Zhebrak A, Vetrov D *et al.* Entangled conditional adversarial autoencoder for *de novo* drug discovery. *Mol. Pharm.* 15(10), 4398–4405 (2018).

16. Grisoni F, Neuhaus CS, Gabernet G, Muller AT, Hiss JA, Schneider G. Designing anticancer peptides by constructive machine learning. *ChemMedChem* 13(13), 1300–1302 (2018).

17. Schneider G. Automating drug discovery. *Nat. Rev. Drug Discov.* 17(2), 97–113 (2018).

18. Sellwood MA, Ahmed M, Segler MH, Brown N. Artificial intelligence in drug discovery. *Future Med. Chem.* 10(17), 2025–2028 (2018).

19. Granda JM, Donina L, Dragone V, Long DL, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* 559(7714), 377–381 (2018).

20. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555(7698), 604–610 (2018).