Editorial

Zhengwei Xie* and Hao Li*

Drug development accelerated by artificial intelligence

https://doi.org/10.1515/mr-2023-0024

The development of biopharmaceuticals, driven by advances in molecular cell biology, has accelerated in the past few decades. However, this development presents a paradox, namely, that highly developed biotechnology has led to lower success rates in clinical trials and higher costs in developing individual drugs. This huge obstacle has made drug development the exclusive privilege of large companies and has limited opportunities for small companies to develop advanced drugs. There is also a huge difference between developed and developing countries in terms of their capacity for pharmaceutical research and development (R&D), and high barriers limit the development of innovative drugs in developing countries.

Since its inception, artificial intelligence (AI)-accelerated drug development has been viewed as a promising alternative to conventional pharmaceutical R&D, with the promise of significantly reducing costs, improving the success, and lowering the barriers for ordinary researchers to develop new drugs.

After more than a decade of AI-based drug development, the technology has moved the field of drug R&D closer to the goal of greater clinical success at lower cost. AI pharmaceutical development has already met several milestones, one of which is deep learning encoding of small molecules. Although small molecules have graph-like structures, they can be unfolded into tree-like structures by breaking the cycles (and adding numerical ring closure labels to show connectivity between non-adjacent atoms), which can be further encoded into one-dimensional strings through tree traversal (SMILES). Machine learning and natural language processing—including recurrent neural networks and recently developed transformer techniques—can be applied to these one-dimensional strings to solve complex problems. Graph neural networks, although seemingly a reasonable approach to representing the structure of molecules, consume excessive amounts of memory and are thus somewhat limited in terms of simulating larger molecules or performing high-throughput calculations.

Representing and predicting the architecture of small molecules are necessary steps before further analysis can be carried out; in particular, the design, modification, and optimization of small molecules for specialized purposes. For example, based on the transformer representation vector, accurate predictions can be made regarding the chemical properties and toxicities of small molecules—such predictions were not previously achievable. Before this kind of R&D can be broadly applied, it is necessary to update the conventional standards for prediction. For example, the prediction accuracy of clinical toxicity is as high as 95 % [1]. The transformer architecture reduces the size of the dataset required to improve the prediction accuracy. The prediction of the one-step inverse synthesis of compounds can also be easily achieved using transformer and Monte Carlo Tree Search [2]; such an approach is simple, elegant, and highly accurate.

Compared to the simplicity of small molecule representation, the deep learning representation of target protein structures is very challenging. This is because of the considerable size of proteins (reflected in their molecular weight) and the fact that the information that would be useful for drug design is derived from local three-dimensional structures and even from the dynamic conformation of the protein. Based on the invariant geometric coding information of local pockets, different design models can be trained on enhanced datasets. The challenge of design is that we need to consider not only high affinity but also other properties, such as ease of synthesis and safety. In short, an elegant and efficient solution remains elusive. A framework for multi-objective design appears to be very important. The serial uses of multiple filter models to screen different properties is essentially brute-force screening and is not sufficiently intelligent. The transformer framework is

***Corresponding authors: Zhengwei Xie**, PKU International Cancer Institute and Departmaent of Pharmacology, School of Basic Medicine, Peking University, Beijing 100871, China; and Peking University - Yunnan Baiyao International Medical Research Center, Peking University Health Science Center, Peking University, Beijing 100191, China, E-mail: xiezhengwei@hsc.pku.edu.cn. https://orcid.org/0000-0001-9572-878X; and **Hao Li**, Department of Biochemistry and Biophysics, University of California San Francisco, 1700 4th Street, San Francisco, CA 94143, USA, E-mail: haoli@genome.ucsf.edu

very powerful but lacks a latent space in which random walks are performed in order to generate systematic variations. The use of variational autoencoders may be a good choice, and there are case studies of this approach [3]. Of course, many factors need to be considered, such as the number of molecules included in the training set, to fully reflect the real neighbors around each molecule.

The advantage of the deep neural network approach is that any complex internal mechanism may be captured in a black box, provided there is a large amount of reliable input-output data. Therefore, even in cases where the target is not well defined, deep neural networks can be used to fit the structures of small molecules and the expression profile changes in the cell caused by the treatment. By connecting expression profile changes related to diseases, the ability of small molecules to regulate key disease-related genes can be predicted, thereby directly predicting drug efficacy [4]. This was not possible before the advent of AI techniques applied in pharmaceutical R&D.

AlphaFold [5], an AI program that predicts protein structures, solves another long-standing problem; that is, how the one-dimensional sequence of amino acids determines the three-dimensional structure. Although AlphaFold is based on a phenomenological approach (rather than first principles), it represents significant progress in AI applications and has practical value. By exchanging the positions of the input amino acid sequences and model parameters, known structures can be input for protein sequence design, which greatly reduces the cost and threshold of molecule design and improves efficiency. In terms of small-molecule drug development, the protein structures predicted by AlphaFold may not be sufficiently accurate at this stage; more systematic evaluations are needed before the utility of this program can be assessed.

Research funding: This work was supported by National Key R&D Program of China (2018YFA0900200) and NSFC (31771519).

Conflict of interest: The authors declare that there is no conflict of interest.

References

1. Honda S, Shi S, Ueda HR. SMILES transformer: pre-trained molecular fingerprint for low data drug discovery. 2019:04738. https://arxiv.org/abs/1911 [Accessed 8 Jun 2023].
2. Lin K, Xu Y, Pei J, Lai L. Automatic retrosynthetic route planning using template-free models. Chem Sci 2020;11:3355–64.
3. Hoffman SC, Chenthamarakshan V, Wadhawan K, Chen P-Y, Das P. Optimizing molecules using efficient queries from property evaluations. Nat Mach Intell 2022;4:21–31.
4. Zhu J, Wang J, Wang X, Gao M, Guo B, Gao M, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. Nat Biotechnol 2021;39:1444–52.
5. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.