A Review on Applications of Machine Learning in Preclinical Drug Discovery

By Anne Liao, Concordia University

Introduction

Drug discovery is a complex and costly venture that involves multiple stages. Scientists first identify potential targets through cellular and genetic evaluation, then proceed with preclinical tests, and finally select patients for clinical trials. Failure in any of these stages results in significant delays or even abandonment of the development. With an estimated expense of 2.6 billion USD and a delivery timeline of over 12 years [1], pharmaceutical researchers are urging to reduce the expenses and expedite the process.

Recently, machine learning (ML), defined by algorithms that are programmed to learn and improve without the need for human intervention, has been increasingly used in various stages of pharmaceutical development. There is unexplored potential for ML in drug discovery that could be applied to areas such as predicting the properties of small molecules, generating new drug design ideas, and reducing costs and timelines to get a drug on the market [2]. This review explores the potential of using ML in the aforementioned areas and the recent advances in preclinical research.

Applications in preclinical drug discovery

Compound screening and optimization

The compound screening and lead discovery stage of drug design occurs after a druggable target protein (linked to an illness) has been identified. The main objective is to identify a cause-effect relationship between a drug candidate and the illness being treated. The process takes around five years to complete, with medicinal chemists synthesizing up to 10,000 compounds during this time [2]. The timeline can be extended or indefinitely delayed if no relationship is found. Additionally, compound screening accounts for roughly 25% of R&D expenditures of a project [3]. Given the high costs and risks of this stage, computational resources, specifically ML, has gathered a lot of interest from pharmaceutical companies in the recent five years.

Traditional optimization technology

The first major usage of computational resources in the screening process involved utilizing supercomputers to perform optimization calculations based on quantum theory [4]. The algorithms are able to calculate the most probable 3D orientation of a small-molecule (a drug candidate) when it interacts with the target protein [5]. The process, referred to as virtual screening, was a major improvement and allowed scientists to visualize microscopic chemical interactions. Traditionally, interactions would be identified using biological assays. However, assays can only give binary results – either there was target engagement or there was not. Using virtual docking methods, a researcher would be able make virtual design changes to a molecule and quantify the resulting effect. Virtual screening is effective in reducing the time spent on chemical synthesis, as compounds with unattractive docking results are immediately abandoned [6].

However, the optimization method comes with many disadvantages. Since optimization algorithms are based on theoretical quantum formulas, a quantum chemist is required to manage these systems. Improving the algorithms require extensive mathematical and physical chemistry considerations, making updates to a software program rare. As well, it is typical that an optimization could take up to a week to generate an accurate 3D structure prediction [7]. With these drawbacks, the cost and time savings of implementing optimization algorithms remains debatable.

Machine learning applications

There is no foolproof way to design a drug and it is common that the rationalization for a particular design change is based on experience and intuition. Consequently, the popularization of ML in image and pattern recognition gained immediate attention from the pharmaceutical industry. Intuition could not be hardcoded into the optimization model; however, with ML, it is possible to present an algorithm with a large amount of data (referred to as the training set) and allow the program to build intuition about the data [8]. The algorithm ~~will be~~ is trained to associate an input with a known output and then be used to predict an outcome given any input.

A supervised learning algorithm can take a small-molecule design as input and generate modifications of that molecule, while maintaining binding and physicochemical properties constant [9]. Using the output of the algorithm, researchers would have a filtered list of potentially viable drug candidates. ML algorithms can be trained on over 700,000 compounds and generate a reliable result in a day [2]. Previously, design modifications would have relied on a researcher's experience and intuition and tested by a docking method over multiple weeks.

With supervised learning, it is essential for the training set to be accurate. Inaccuracies will be amplified in the algorithm and result in incorrect conclusions [10]. Thus, the capability of supervised learning is limited, as it will not generate innovative ideas and cannot perform better than the scientists that filtered the data and trained the algorithm [8]. These weaknesses led to the implementation of a more sophisticated ML design, known as unsupervised learning.

In unsupervised learning, an algorithm is trained on unlabelled data (i.e., data with no drawn conclusions) and searches for patterns and connections among the data. By implementing this method of learning, an algorithm can supersede researcher capabilities and generate innovative drug design ideas [11]. In 2021, a 3D ML-based modelling algorithm developed by Ganea et al. from MIT is shown to out-perform other mainstream open-source options [12]. However, the model does not perform consistently across various chemical groups due to limited datasets. If more data was made public, it would be possible for experienced computer scientists and medicinal researchers across the globe to collaborate on an open-source ML algorithm. Presently, more advancements will be required to implement a reliable unsupervised algorithm, and data accessibility and quality continue to be the largest barrier.

Larger pharmaceutical companies remain the sole pioneers to the advancement of ML technology. With their extensive high-quality experimental data, the barrier for entry is lower and the cost and time savings are compelling. The development of in-house algorithms also allows computationally designed drugs to maintain a high market value due to their proprietary technologies. In 2021, Exscientia, a pharmatech leading the AI driven drug design movement, became the first company to have two completely AI-designed molecules enter Phase I of clinical

trials to treat advanced tumours [13]. Shortly afterwards, in the last quarter of 2021, Insilico Medicine also nominated an AI-designed drug candidate to clinical trials for treatment of idiopathic pulmonary fibrosis (IPF) and is currently undergoing in-human studies [14]. Both these companies managed to shorten the preclinical timeline (typically 5 years) to less than 18 months and reduce preclinical costs to less than 5 million USD [15].

Conclusion

The aim of the present review was to examine the current usage of machine learning (ML) in drug discovery and explore the advantages and disadvantages of using such technologies. This paper has shown that there are various areas in drug discovery where ML has been successfully applied to, for example, in screening and drug design. As discussed, the usage of ML methods has dramatically improved the efficiency of modelling, generated innovative drug design ideas, and reduced costs and timelines compared to conventional methods such as optimization modelling. There are still several areas where ML needs improvement, such as data accessibility and unexplored chemical space. However, the chemical intuition that a computer algorithm can generate without the supervision of humans is a substantial feat. Overall, this review strengthens the idea that ML is becoming a dominant force in drug discovery and pursuing further development in this area would be beneficial for the medical field.

References

[1] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan, "Advancing Drug Discovery via Artificial Intelligence," *Trends in Pharmacological Sciences*, vol. 40, no. 8, pp. 592-604, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.tips.2019.06.004.

[2] J. Vamathevan *et al.*, "Applications of machine learning in drug discovery and development," *Nat. Rev. Drug Discovery*, vol. 18, no. 6, pp. 463-477, 2019, doi: 10.1038/s41573-019-0024-5.

[3] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: A systematic review," *Health Policy*, vol. 100, no. 1, pp. 4-17, 2011/04/01/ 2011, doi: https://doi.org/10.1016/j.healthpol.2010.12.002.

[4] O. A. von Lilienfeld, "Quantum Machine Learning in Chemical Compound Space," *Angewandte Chemie International Edition*, https://doi.org/10.1002/anie.201709686 vol. 57, no. 16, pp. 4164-4169, 2018/04/09 2018, doi: https://doi.org/10.1002/anie.201709686.

[5] X. Liu, A. P. Ijzerman, and G. J. P. van Westen, "Computational Approaches for De Novo Drug Design: Past, Present, and Future," in *Artificial Neural Networks*, H. Cartwright Ed. New York, NY: Springer US, 2021, pp. 139-165.

[6] X. Lin, X. Li, and X. Lin, "A Review on Applications of Computational Methods in Drug Screening and Design," *Molecules*, vol. 25, no. 6, p. 1375, 2020. [Online]. Available: https://www.mdpi.com/1420-3049/25/6/1375.

[7] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature reviews Drug discovery*, vol. 3, no. 11, pp. 935-949, 2004.

[8] R. S. K. Vijayan, J. Kihlberg, J. B. Cross, and V. Poongavanam, "Enhancing preclinical drug discovery with artificial intelligence," *Drug Discovery Today*, 2021/11/25/ 2021, doi: https://doi.org/10.1016/j.drudis.2021.11.023.

[9] F. Miljković, R. Rodríguez-Pérez, and J. Bajorath, "Impact of Artificial Intelligence on Compound Discovery, Design, and Synthesis," *ACS Omega*, vol. 6, no. 49, pp. 33293-33299, 2021/12/14 2021, doi: 10.1021/acsomega.1c05512.

[10] L. H. Mervin, S. Johansson, E. Semenova, K. A. Giblin, and O. Engkvist, "Uncertainty quantification in drug design," *Drug Discovery Today*, vol. 26, no. 2, pp. 474-489, 2021/02/01/ 2021, doi: https://doi.org/10.1016/j.drudis.2020.11.027.

[11] T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, and S. K. Saikin, "Autonomous Molecular Design: Then and Now," *ACS Applied Materials & Interfaces*, vol. 11, no. 28, pp. 24825-24836, 2019/07/17 2019, doi: 10.1021/acsami.9b01226.

[12] O.-E. Ganea *et al.*, "GEOMOL: torsional geometric generation of molecular 3D conformer ensembles," *arXiv.org, e-Print Arch., Phys.*, pp. 1-25, 2021. [Online]. Available: http://arxiv.org/archive/physics.

[13] "Exscientia Announces First AI-Designed Immuno-Oncology Drug to Enter Clinical Trials." http://www.drugdiscoverytoday.com/view/47868/exscientia-announces-first-ai-designed-immuno-oncology-drug-to-enter-clinical-trials/ (accessed 9 March 2022).

[14] M. Colangelo. "Insilico Medicine Initiates Trial – Doses First Human With AI-Discovered Drug." https://www.aitimejournal.com/@margaretta.colangelo/insilico-medicine-starts-trial-and-doses-first-human-with-ai-discovered-drug#:~:text=Insilico%20Medicine%20Initiates%20Trial%20%E2%80%93%20Doses%20First%20Human%20With%20AI%2DDiscovered%20Drug&text=Insilico%20Medicine

%20announced%20today%20that,drug%20candidate%20for%20pulmonary%20fibrosis. (accessed 9 March 2022).

[15] "Bridging Biology, Chemistry and Medicine in Drug Discovery and Development with End-to-End Artificial Intelligence." https://insilico.com/blog/fih (accessed 9 March 2022).