Commentary

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

Future Medicinal Chemistry

Artificial intelligence in drug discovery

Matthew A Sellwood[1], Mohamed Ahmed[1], Marwin HS Segler[1] & Nathan Brown*,[1]
[1]BenevolentAI, 40 Churchway, London, NW1 1LW, UK
*Author for correspondence: nathan.brown@benevolent.ai

"While many of the new approaches have yet to bear fruit in terms of drugs being progressed to market, initial reports tend toward the belief that they will become even more integral in the drug discovery process than has hitherto been seen."

First draft submitted: 1 June 2018; Accepted for publication: 5 June 2018; Published online: 13 August 2018

What is artificial intelligence & machine learning?

Artificial intelligence (AI), and the subfield of machine learning (ML), study the processes and practicalities of enabling machines to skilfully perform intelligent tasks, without explicitly being programmed for those tasks. Recently, AI systems have neared or surpassed human performance in several tasks, such as game playing and image recognition [1], but these have typically been quite narrow and focused domains. Nonetheless, AI in its various forms is today successfully applied across a large range of domains and for challenging tasks, ranging from robotics, speech translation, image analysis and logistics to its ongoing use in designing molecules.

Since the 1960s, medicinal chemistry has applied AI in various forms and with varying degrees of success to the design compounds. Supervised learning, where labeled training datasets are used to train models is extensively applied. An example is the quantitative structure–activity relationship (QSAR) approach, which is widely used to predict properties, such as logP, solubility and bioactivity, for given chemical structures. Conversely, unsupervised learning, which does not rely on labels, is also popular in medicinal chemistry, with examples such as hierarchical clustering, algorithms and principal components analysis being used extensively to analyze and break down large molecular libraries into smaller collections of similar compounds [2].

Hype versus hope: managing expectations

The ultimate goals of applying AI and ML methods to challenges in drug discovery remain the same as they ever were: bringing the best drugs to the clinic to satisfy unmet medical need. For drug discovery and medicinal chemistry specifically, this involves tasks in identifying drug targets, identifying lead compounds, optimizing their designs against multiple property profiles of interest and identifying synthetic routes to realize the composition of matter.

AI is often seen as a magic button that can be pressed at will to produce the perfect output, often regardless of input. Whether the AI challenge is to design the perfect image of a cat from a model trained on images of cats, a car that is able to drive itself without making a single mistake, or a drug that can be designed to treat a disease safely and efficaciously. While AI is not the answer to every challenge, it is a useful tool that if used correctly can help to augment current understanding and drive new discoveries. Within medicinal chemistry and drug discovery, the best AI is not necessarily a single AI that can autonomously design a new drug, but one or many different AIs, that enable better understanding and the design of new inputs, throughout the drug discovery process from target selection, hit identification, lead optimization to preclinical studies and clinical trials.

Molecular design

One of the fundamental questions one can ask in drug discovery is: which chemical structures will elicit the desired property profile. *De novo* molecular design can combine optimization parameters such as predictive models and molecular similarity, with molecule generation and search to simulate design–make–test cycles [3]. These *in silico* design loops then provide a list of candidate solutions that identify chemical structures that are predicted to be optimal for the profile defined. However, significant challenges remain with regard to the synthetic tractability of these candidates.

newlands press

An approach to molecular design published recently [4], applies analogs of evolution to optimize chemical structures against a defined set of objectives, such that a structure with the desired profile emerges, known as multiparameter optimization [5]. The multiobjective automated replacement of fragments algorithm proceeds by initializing a population of candidate structures, which are iteratively evaluated, sampled and scored to optimize against the structure profile of interest. The multiobjective automated replacement of fragments algorithm uses a database of derived building blocks from known synthetic organic chemistry, called synthetic disconnection rules, where the bonding patterns and frequency of occurrence of each, are retained. Replacement substructures are selected using a new algorithm called rapid alignment of topological structures to simultaneously balance the exploration of the replacements while minimizing the disruption of the information contained in the candidate structures. This approach was demonstrated to optimize the potency of a CDK2 inhibitor, while also improving its cell permeability. Furthermore, due to the approach used to generate the list of molecular building blocks, synthetic accessibility is indirectly considered, but by no means is this measure of synthetic accessibility appropriate in all cases.

One way that challenges in the automated design of compounds of synthetic tractability has been tackled is using models based on synthetic rules, which combine building blocks using standard synthetic couplings [6]. However, these approaches tend to limit the exploration of the relevant chemical space [7]. An alternative way to generate new chemical structures has recently been proposed by Gomez-Bombarelli *et al.* [8] and Segler *et al.* [9], these approaches introduce AI-based generative models for molecules. The models are trained on large datasets of molecular structures from exemplified medicinal chemistry space, for example, ChEMBL. These generative models learn a distribution over the molecules in the dataset. From this distribution, these approaches permit the sampling of novel molecules from the chemistry space that has been learned to be more 'drug-like'. Recently, a number of neural generative methods have been proposed and benchmarked for molecular design, with recent work concluding that recurrent neural networks currently perform the best [10]. However, the main challenge of synthetic accessibility remains with further work in the field required.

The current active landscape of research in the area of automated molecular design suggests that no one solution is appropriate for all applications. Recent advances in synthetic tractability (*vide infra*) will undoubtedly assist in this task, additionally improved exploration and exploitation of the relevant chemical space remains a significant obstacle to be able to home in on those chemical structures most relevant to progress to synthesis and testing. One particular challenge in this arena is the ability to predict reliable properties, such as biological activity.

Predictive modeling

From the origins of atomistic theory, chemists have endeavored to predict the properties of compounds without requiring to synthesize these compounds. Alexander Crum Brown stated in 1869, that physiological response of a compound is merely a function of its chemical constitution, however defining that function remains challenging. QSARs and its relations were first proposed by Hansch and Fujita in 1962, and since this time they have remained an active area of research. The work on QSAR has led to advances into the routine of particular physicochemical property predictions, notably exemplified by ClogP, for calculating the octanol/water partition coefficient [11].

Since the formal advent of QSAR over 50 years ago, the numbers of modeling techniques, representations of molecules and volume of data and compute resource available have increased significantly. The advances in all of these fields mean that techniques such as deep learning that previously were not appropriate or available to these datasets can now be utilized. We now have access to large quantities of chemical structure data together with measured end points of relevance, from which it is possible to generate predictive models. However, there still remains a limited quantity of these data and even when access is available, the quality of highly variable. Here, the expectation is that more modern ML methods will be able to tackle these noisy data.

One of the first applications of deep learning to chemical property prediction was as a result of the Merck molecular activity challenge, with multitask neural networks to predict not only one end point, but multiple end points simultaneously [12]. Deep learning chemical property prediction is now a very active area of research [13].

Synthesis planning

Planning the synthesis of novel compounds requires expertise, experience and creativity. Even though chemists can now synthesize almost everything they so desire, some compounds present themselves as tough nuts to crack. In addition, *de novo* design can easily suggest millions of chemical structures, only offering reasons why they should be

made and not how they can be realized. Computer-aided synthesis planning (CASP) can help in both situations: by providing alternative routes or helping to prioritize compounds which can be readily synthesized.

CASP has a long tradition, starting in the 1960s [14,15]. Ironically however, the main concept developed for CASP, working backward from the target using transformation rules and heuristics, which is now known as retrosynthetic analysis, turned out to be tremendously helpful for humans, but less so for machines.

Recently, however, principled headway has been made. Grzybowski and coworkers reinvigorated the classic idea of heuristic-based analysis by letting experts code a large number of rules into the machine and demonstrated that the machine was able to propose tractable routes for eight medicinally relevant compounds [16].

Going further, Segler *et al.* demonstrated that the computer can even learn the rules of organic chemistry autonomously from chemical reaction data without expert input [17]. Using deep neural networks they first let the machine learn to focus on the most promising rules for retroanalysis, which are then submitted to reaction prediction in combination with a modern Monte-Carlo tree search algorithm. A double-blind study, synthetic organic chemists on an average, considered the routes generated by this method to be at par with routes taken from the literature.

Feedback loop

Medicinal chemistry and drug discovery projects operate as feedback loops, exemplified as the classical 'design–make–test' cycle, where compounds that are designed must be synthesized and tested experimentally to provide feedback for further decision making. Evidently, this process is relatively slow and expensive. It may take weeks to generate experimental data from which new design decisions can be made. Using methods described above in the 'Molecular design' section to generate candidate solutions with appropriate profiles and even how to make the compounds, will undoubtedly streamline this process. However, what if even further improvement could be made.

Active learning is an area of ML [18] where decisions on the next data point to be – labeled or compound to be synthesized and – tested can be made effectively and efficiently. One of the expected strengths of this approach is to be able to simultaneously make predictions for compounds that will progress a project, but also more rapidly identify the compounds that should be synthesized to improve the models. Such improvement in the models can thereby indirectly improve and streamline the drug discovery process as the models will improve in prediction of quality much more rapidly.

While some scientific efforts have been made in the area of active learning in drug discovery, it remains an area that requires significant amount of investment to demonstrate its worth prospectively to commit to make and test the identified compounds [19]. It is challenging to elicit confidence from experimentalists to make compounds that will not necessarily meet the current objectives of a drug discovery program, but will likely improve the process going forward. As such, this is an example of AI and ML that is not only bound by its direct importance to drug discovery, but also the support from those scientists who will work closely with these systems and need to make and test the compounds as we increasingly automate certain aspects of drug discovery, while ensuring that humans continue to be heavily involved in the process [20].

Conclusion & future perspective

Recent advances in AI and ML have returned these methods and approaches from their wilderness years. While many of the new approaches have yet to bear fruit in terms of drugs being progressed to market, initial reports tend toward the belief that they will become even more integral in the drug discovery process than has hitherto been seen. Through applications of new and promising techniques, it has been shown that the new systems can design new chemical structures effectively, predicted for the desired molecular property profiles and even how to synthesize those compounds. While many of these areas of research have been promised many times before, it is becoming a perfect storm of many different advances simultaneously reaching their apogee.

Financial & competing interests disclosure

The authors have no 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.

No writing assistance was utilized in the production of this manuscript.

References

1. Mnih V, Kavukcuoglu K, Silver D *et al.* Human-level control through deep reinforcement learning. *Nature* 518(7540), 529–533 (2015).
2. Butina D. Unsupervised database clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large datasets. *J. Chem. Inf. Comput. Sci.* 39(4), 747–750 (1999).
3. Brown N. *In Silico Medicinal Chemistry: Computational Methods to Support Drug Design*. Royal Society of Chemistry, Cambridge, UK (2016)
4. Firth NC, Atrash B, Brown N, Blagg J. MOARF, an integrated workflow for multiobjective optimization: implementation, synthesis, and biological evaluation. *J. Chem. Inf. Model.* 55(6), 1169–1180 (2015).
5. Brown N, McKay B, Gilardoni F, Gasteiger J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* 44(3), 1079–1087 (2004).
6. Hartenfeller M, Zettl H, Walter M *et al.* DOGS: reaction-driven *de novo* design of bioactive compounds. *PLoS Comput. Biol.* 8(2), e1002380 (2012).
7. Gillet VJ, Bodkin MJ, Hristozov D. Multiobjective *de novo* design of synthetically accessible compounds. In: *De Novo Molecular Design*. Schneider G (Ed.). Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany, 267–285 (2013)
8. Gómez-Bombarelli R, Wei JN, Duvenaud D *et al.* DOGS: reaction-driven de novo design of bioactive compounds. *ACS Cent. Sci.* 4(2), 268–276 (2018).
9. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4(1), 120–131 (2018).
10. Neil D, Segler M, Guasch L *et al.* Exploring deep recurrent models with reinforcement learning for molecule design. (2018). https://openreview.net/pdf?id=Bk0xiI1Dz
11. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194, 178 (1962).
12. Dahl GE, Jaitly N, Salakhutdinov R. Multitask neural networks for QSAR predictions. arXiv[stat.ML]. (2014). http://arxiv.org/abs/1406.1231
13. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. *J. Comput. Chem.* 38(16), 1291–1307 (2017).
14. Vléduts GÉ. Concerning one system of classification and codification of organic reactions. *Inform. Stor. Retr.* 1(2), 117–146 (1963).
15. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. *Science* 166(3902), 178–192 (1969).
16. Klucznik T, Mikulak-Klucznik B, McCormack MP *et al.* Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* 4(3), 522–532 (2018).
17. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555(7698), 604–610 (2018).
18. Cohn DA, Ghahramani Z, Jordan MI. Active learning with statistical models. In: *Advances in Neural Information Processing Systems 7*. Tesauro G, Touretzky DS, Leen TK (Eds). The MIT Press, Cambridge, MA, USA, 705–712 (1995).
19. Reker D, Schneider G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today* 20(4), 458–465 (2015).
20. Schneider G. Automating drug discovery. *Nat. Rev. Drug Discov.* 17(2), 97–113 (2018).