Artificial Intelligence in the Life Sciences 2 (2022) 100041

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Viewpoint

AI for drug design: From explicit rules to deep learning

Lewis Mervin [a,*], Samuel Genheden [b], Ola Engkvist [b]

[a] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[b] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

In this viewpoint we will briefly discuss the historical development of applying machine learning (ML) and artificial intelligence (AI) to drug design. The goal is to try to describe what is different now compared to earlier periods of enthusiasm for applying AI to drug design. The term AI was coined in the 1950's in a period of strong excitement for the promise of AI. Several more periods of positivity followed, for instance for the rule-based expert systems in the 1980's, as well as several so-called "AI-winters" with disillusion about the possibilities with AI. However, the continuous innovation in algorithms like backpropagation and convolutional neural networks (CNNs), together with Moore's law providing exponentially faster hardware, made it (in hindsight) unavoidable that ML and AI would make a largescale comeback and impact many parts of society [1]. Whilst CNNs are just one important deep learning (DL) architecture, others include recurrent neural networks (RNNs), variational auto-encoders (VAEs), multi-layer feed forward networks (FFNs) and transformers. One notable inflexion point was the success of a CNN in the ImageNet challenge in 2012, where the winning contribution made a marked improvement in classification accuracy for different image types [2]. This success has been followed by many more, for example AlphaGo, AlphaZero and AlphaFold2. Such progress was made possible through a virtuous cycle of faster hardware and improved, more flexible algorithms that can better leverage the available data. We argue that image classification and language translation have both seen a movement over time, away from rule-based systems to more flexible systems based on DL and large training sets.

The use of cheminformatics applications in drug design has also experienced similar peaks and troughs. An example of such a peak was in the late 1990's, with the enthusiasm to combine ML/AI with combinatorial chemistry and high-throughput screening. We would like in this viewpoint to argue that ML/AI for drug design is making the same journey as ML/AI in general, to move away from rule-based models to more flexible deep learning-based systems, as illustrated in Fig. 1. This is an on-going journey, and so far, not all of the deep learning-based methods have shown superior performance for cheminformatics applications.

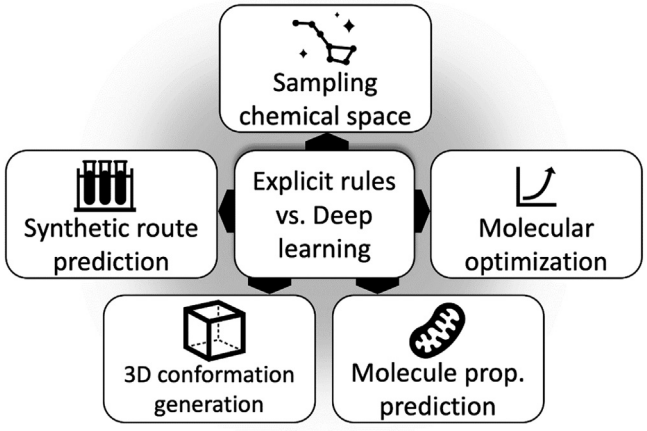

Fig. 1. An overview of the tasks impacted by the choice of explicit rule- versus deep learning-based systems. We see that cheminformatics is experiencing a general movement away from rule-based models to more flexible deep learning-based systems (a trend highlighted via the central cell in the diagram). The impact of this is particularly evident for tasks such as sampling of the chemical space, molecular optimization, property prediction, 3D conformer generation and synthetic route prediction, as highlighted via the outer boxes.

1. Sampling of the chemical space

The estimated size of the relevant druglike chemical is 10^{60}. This is evidently so large that it can't be enumerated, and other methods must be used to sample it. Efficient sampling of the whole chemical space is particularly necessary for applications such as "scaffold-hopping" where a large part of the molecular structure requires replacement due a liability, such as metabolic stability. Around the turn of the century genetic algorithms (GA) were developed to sample the chemical space, given a set of building blocks [3]. While some interesting proof-of-principle studies were published, these rule-based methods failed to gain traction in practice. One issue was the synthetic accessibility of the generated molecules, as well as blindly merging reasonabe-looking building blocks, which did not necessarily produce reasonabe-looking molecules. In contrast, deep learning-based methods that are capable of sampling the chemical space have been available since 2016, where one popular algorithm used is the RNN [4]. An RNN is trained on a set of molecules and will generate molecules that were not observed in the training set, but rather generated from the same area of the chemical space. By comparing outputs with the GDB-13 database [5], it was shown an RNN trained on a fraction of GDB-13 will cover most of the database, whilst only a small fraction of the generated molecules do not form part of the database [6]. The RNN mimics the properties of the molecular database

* Corresponding author.
E-mail address: lewis.mervin1@astrazeneca.com (L. Mervin).

https://doi.org/10.1016/j.ailsci.2022.100041
Received 6 July 2022; Accepted 10 July 2022
Available online 12 July 2022

it has been trained on. For instance, if it is trained on easy to synthesize compounds, then the generated molecules will be easy to synthesize; conversely, an RNN trained on difficult to synthesize natural products will generate molecules more difficult to synthesize. Hence, the key difference between generating molecules with an RNN and a GA is that an RNN generates molecules from a probability distribution learned from the training set, so it is highly unlikely to generate molecules with completely different properties, such as synthetic accessibility. In comparison all combinations of building blocks are generated with the same probability when using a GA and thus the generated molecules are not directed towards similar properties.

2. Molecular optimization

In molecular optimization, a starting molecule is optimized to improve certain properties such as potency, selectivity, solubility etc., and the structural modifications are usually rather small. Optimization has historically concentrated on some type of enumeration, such as matched molecular pairs (MMPs) [7]. MMPs are an intuitive way of modifying a molecule and can provide an understanding how structural modifications influence molecular properties. However, this rule-based MMP method treats alltransformations with the same probability. One DL-based alternative is to train a transformer on a set of historical transformations, and then use the trained transformer network to identify suitable transformations. Indeed, transformers trained on MMPs pairs were able to identify a larger number of suitable molecular transformations than compared with asimple MMP enumeration [8].

3. Synthetic route prediction

Rule-based synthetic route prediction has its roots in the work of EJ Corey in the 1960′s and there were no popular alternatives to rule-based methods until relatively recently [9]. A main reason for the lack of progress was the availability of reaction data sets. This changed in recent years due to a reaction set derived from US patent office records, and the access granted to academic groups for proprietary reaction data from the scientific literature. The access to large reaction data sets triggered the development of several synthetic route prediction tools based on deep learning. Such tools can generally be divided into template-based and template-free methods. Template-based methods consist of deep learning-based forward and backward neural networks combined with efficient tree search algorithms to explore the space of potential synthetic routes [10]. Template-free methods usually predict forward or backward single reaction steps and are most commonly based on transformers [11]. Unfortunately it is difficult to compare DL-based and rule-based synthetic route prediction tools due to the proprietary nature of most reaction data and state-of-the-art reaction rule sets.

4. Molecular property prediction

Molecular property prediction has utilised ML techniques to model the relationship between the input descriptors and the model endpoint for some time [12]. The descriptors have been historically generated through rule-based methods, such as molecular fingerprints [13]. However, graph-convolutional methods have also been developed [14]. In these methods, the molecular graphs are used as input descriptors for modeling the property of interest. The graph convolutional neural network is trained to determine which part of the molecular graph is relevant for prediction accuracy, and thus provides greater flexibility than molecular fingerprints [15]. With the increase in computational power, multi-task deep learning has further increased in popularity and can also be combined with either graph convolutional methods or molecular fingerprints. Indeed, graph convolutional networks outperformed molecular fingerprint-based methods in several experiments [12]. However, other studies report more equal performances between for the two approaches, so the improvement gains are likely context dependent. Performance benefits for DL-based methods were more often observed in studies with larger data sets, where there is a wider scope for the prediction algorithm to detemine the important features and/or related tasks [8].

5. 3D conformational generation

Conformation generation from a molecular graph is mainly conducted using rule-based methods. These methods have served the cheminformatics community well over the years [16]. The approach includes generating a 3D conformation from a molecular graph, as well as generating all relevant low energy conformations. It is hence logical to investigate end-to-end generative methods, given the development of generative methods for graphs, for example, starting with graph generation and including calculation of the available low energy conformations. Several promising examples based on VAE [17] and flow networks have been recently published [18]. In particular, a torsional diffusion model was shown to outperform rule-based conformational generators [19]. A comprehensive review of deep learning-based 3D molecular generation has also just been published [20].

Conclusions

Above are examples describing how ML/AI for drug design has been evolving during the last few years, from rule-based models to DL-based models. It is oulined that many of the key cheminformatics tasks can now be conducted with DL instead of rule-based methods. Given the fast development of novel, better algorithms it is expected that this development will further continue. However, there is also an unavoidable synergy between rule-based methods and DL, where DL can enhance a traditionally rule-based method. An example is molecular dynamics (MD) simulations, where there is huge potential to improve MD simulations with DL. For example, DL-based Boltzmann generators can be used to speed-up MD simulations [21], whilst DL based force fields might improve accuracy for binding affinity calculations in the future [22]. Thus, it looks like we currently have a paradigm shift in cheminformatics, where the progress in hardware, data availability and novel DL algorithms have together reached a tipping point to impact cheminformatics. One should proceed with caution, however, since deep learning-based methods also have well-known drawbacks, as highlighted by Gary Marcus and others [23]. For example, DL excels in pattern recognition but lacks the capability of reasoning. An active promising research field is therefore to combine DL with neuro-symbolic reasoning [24], which takes advantage of both the DL-based propensity for pattern recognition and the capacitiy of reasoning from neuro-symbolic systems.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[2] Krizhevsky A, Sutskever I, Hinton GE. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Nevada: Curran Associates Inc.: Lake Tahoe; 2012. p. 1097–105.
[3] Schneider G, Lee M-L, Stahl M, Schneider P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. J Comput-Aided Mol Des 2000;14:487–94.
[4] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 2018;4:120–31.
[5] Blum LC, Reymond J-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 2009;131:8732–3.
[6] Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O. Randomized SMILES strings improve the quality of molecular generative models. J Cheminform 2019;11:1–13.
[7] Griffen E, Leach AG, Robb GR, Warner DJ. Matched molecular pairs as a medicinal chemistry tool. J Med Chem 2011;54:7739–50.
[8] He J, You H, Sandström E, Nittinger E, Bjerrum EJ, Tyrchan C, Czechtizky W, Engkvist O. Molecular optimization by capturing chemist's intuition using deep neural networks. J Cheminform 2021;13:26.
[9] Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. Acc Chem Res 2018;51:1281–9.
[10] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555:604–10.

[11] Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5:1572–83.
[12] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A, Isayev O, Curtalolo S, Fourches D, Cohen Y, Aspuru-Guzik A, Winkler DA, Agrafiotis D, Cherkasov A, Tropsha A. QSAR without borders. Chem Soc Rev 2020.
[13] Sydow D, Burggraaff L, Szengel A, van Vlijmen HWT, AP IJ, van Westen GJP, Volkamer A. Advances and challenges in computational target prediction. J Chem Inf Model 2019;59:1728–42.
[14] Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput-Aided Mol Des 2016;30:595–608.
[15] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59:3370–88.
[16] Hawkins PCD. Conformation generation: the state of the Art. J Chem Inf Model 2017;57:1747–56.
[17] Kang, S.-g.; Weber, J.K.; Morrone, J.A.; Zhang, L.; Huynh, T.; Cornell, W.D., In-pocket 3D graphs enhance ligand-target compatibility in generative small-molecule creation. *arXiv preprint arXiv:2204.02513* 2022.
[18] Gebauer NWA, Gastegger M, Hessmann SSP, Müller K-R, Schütt KT. Inverse design of 3d molecular structures with conditional generative neural networks. Nat Commun 2022;13:973.
[19] Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; Jaakkola, T., Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729* 2022.
[20] Xie W, Wang F, Li Y, Lai L, Pei J. Advances and challenges in De Novo drug design using three-dimensional deep generative models. J Chem Inf Model 2022;62:2269–79.
[21] Noé F, Olsson S, Köhler J, Wu H. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. Science 2019;365:eaaw1147.
[22] Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, Schütt KT, Tkatchenko A, Müller K-R. Machine learning force fields. Chem Rev 2021;121:10142–86.
[23] Marcus G. Deep learning is hitting a wall. Nautilus, Accessed 2022:03–11.
[24] Corchado, J. Neuro-symbolic reasoning-a solution for complex problemas. In INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS. LONDON, UK, 1995.