Bioorganic & Medicinal Chemistry Letters 28 (2018) 2807–2815

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

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

Digest

Recent applications of machine learning in medicinal chemistry

Jane Panteleev[a,*,1], Hua Gao[a,*,1], Lei Jia[b,*,1]

[a] Amgen Discovery Research, 360 Binney St., Cambridge, MA 02141, USA
[b] Amgen Discovery Research, One Amgen Center Dr., Thousand Oaks, CA 91320, USA

A B S T R A C T

In recent decades, artificial intelligence and machine learning have played a significant role in increasing the efficiency of processes across a wide spectrum of industries. When it comes to the pharmaceutical and biotechnology sectors, numerous tools enabled by advancement of computer science have been developed and are now routinely utilized. However, there are many aspects of the drug discovery process, which can further benefit from refinement of computational methods and tools, as well as improvement of accessibility of these new technologies. In this review, examples of recent developments in machine learning application are described, which have the potential to impact different parts of the drug discovery and development flow scheme. Notably, new deep learning-based approaches across compound design and synthesis, prediction of binding, activity and ADMET properties, as well as applications of genetic algorithms are highlighted.

General concepts in the field of artificial intelligence (AI) increasingly feature in discussions of predictive modeling and optimization of the medicinal chemistry processes in drug discovery. One of the goals of AI is the establishment of machine learning (ML) platforms that allow progressive improvement of model performance. For instance, a recent class of ML algorithms based on complex architectures of neural networks, deep learning (DL), is prominent in the current literature as a possible tool for improving existing models or developing novel computational platforms (Fig. 1). This perspective aims to give an overview of examples showcasing how new ML methods, including DL, are applied in different areas of the drug discovery process.

Recent advances in ML have led to significant expansion and influence of technologies utilizing AI on our everyday lives. Through countless examples of applications empowered by rapid growth of data (big data), ranging from Apple's SIRI and Tesla's self-piloting Model S vehicle to the Google Photo auto classification feature and advanced risk assessment by the credit card issuers, AI has led to significant changes across a broad spectrum of industries. Although application of many of these methods to research and development in pharmaceutical and biotechnology industries is still in the early phase, the potential of AI to revolutionize the drug discovery process has already been demonstrated by the uptake and prevalent use of absorption, distribution, metabolism, excretion, and toxicity (ADMET) predictive tools, virtual screening, and quantitative structure activity relationship (QSAR) modeling.[1] Development and implementation of AI methods could considerably benefit more aspects of early drug discovery. This could include more accurately identifying hits from high throughput screening (HTS) and hit expansion, effectively designing novel drug like compounds and synthetic routes toward them, improving efficiency of lead optimization through design of more accurate ADMET and QSAR models, and lowering the overall cost and time of the drug discovery process (Fig. 2).

Development and application of *in silico* QSAR models to predict drug activity, has become increasingly utilized in drug development and discovery over the past few decades.[2] Simple machine learning algorithms such as multiple linear regression (MLR) and partial least squares (PLS) have been used for model development for relatively small data sets. With the application of HTS assays in drug discovery, large amounts of data on activity of diverse chemical matter have been generated,[3] and more sophisticated machine learning algorithms[i] such as random forests (RF),[4] support vector machines (SVM),[5] artificial neural network (ANN),[6] and Cubist[7] have been developed and used successfully to develop various types of *in silico* ADMET and QSAR predictive models (Fig. 3).

The current era of AI is led by advancement of deep learning (DL) technology and the methods around its implementation, which have demonstrated advantages in many areas including drug discovery.[8–10] DL expands on the application of artificial neural networks in ML,

* Corresponding authors.
E-mail address: ljia@amgen.com (L. Jia).
[1] All three authors contributed equally to this article.
[i] While detailed description of existing ML algorithms and statistical learning models falls outside of the scope of this review, numerous publications offer a good introduction to statistical learning principles. For an example see: James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. New York, NY: Springer; 2013.

https://doi.org/10.1016/j.bmcl.2018.06.046
Received 10 April 2018; Received in revised form 24 June 2018; Accepted 26 June 2018
Available online 28 June 2018
0960-894X/

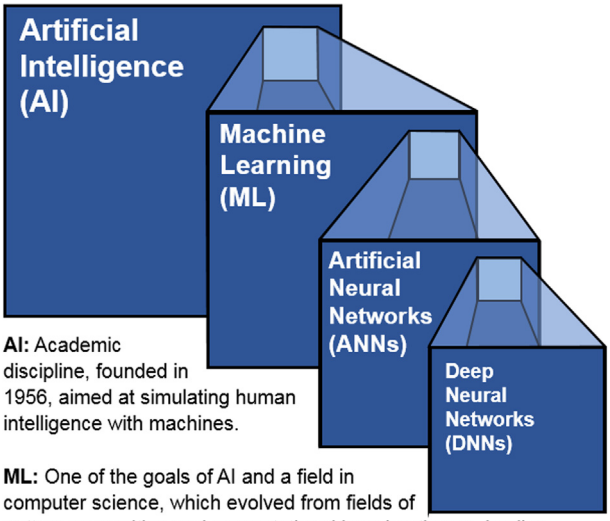

Fig. 1. Key definitions and relationships between artificial intelligence and deep learning.

where more complex network architectures are used (Fig. 4). Typically, multiple fully connected or convolutional hidden layers are utilized and interact in different ways to define platforms such as deep neural networks (DNNs), recurrent neural networks (RNNs), and convolutional deep neural networks (CNNs). DNNs are artificial neural networks with multiple hidden layers, typically with data flow in the forward direction, with capability to model complex non-linear relationships. RNNs can use internal memory to process arbitrary sequences of inputs, with data flow in any direction. RNNs are frequently used for text and speech recognition. The key feature of CNNs compared to DNNs is that CNNs can learn to recognize patterns across space and are effectively used in image and pattern recognition.

The use of DL methods has some advantages over traditional ML methods, namely in the capability of handling larger datasets, being able to utilize larger numbers of descriptors (or features), and obviating the need for costly feature engineering (Fig. 3). A predictive modeling and analytics contest (Kaggle competition) hosted by Merck, Inc. in 2012 demonstrated the high performance of deep neural nets, with methods utilizing DNNs winning the contest by predicting binding and ADMET properties of 15 data sets of small molecules.[11] DL-based methods have since been actively developed in the context of drug discovery,[10] especially when building more sophisticated and presumably more robust ADMET models.[12–18] In recent years, several start-up companies, including Accutar Biotechnology, Atomwise, Berg, Exscientia, Insilico, Insili.com, and Numerate, have been founded based on DL or non-DL based AI technology platforms, which are being actively explored in collaboration with pharmaceutical and biotechnology companies.[19]

Although a complete review of machine learning, and its applications in medicinal chemistry would require a more extensive series of publications, herein we focus on selected topics in early drug discovery which have been influenced by the most recent advancements in ML, including examples in structure generation and *de novo* design, synthetic assessment, ligand binding and activity prediction, ADMET predictions, and application of genetic algorithms in medicinal chemistry.

Recent applications of ML in structure generation

Although numerous computational approaches exist to generate and enrich virtual libraries for screening or hit expansion purposes, application of AI methods for the generation of novel chemical structures during lead optimization has been fairly confined to fragment growing or scaffold hopping strategies. Searching of extensive libraries of existing or virtual compounds by using pharmacophore models or virtual docking have been central to Ligand-Based Drug Design (LBDD) and

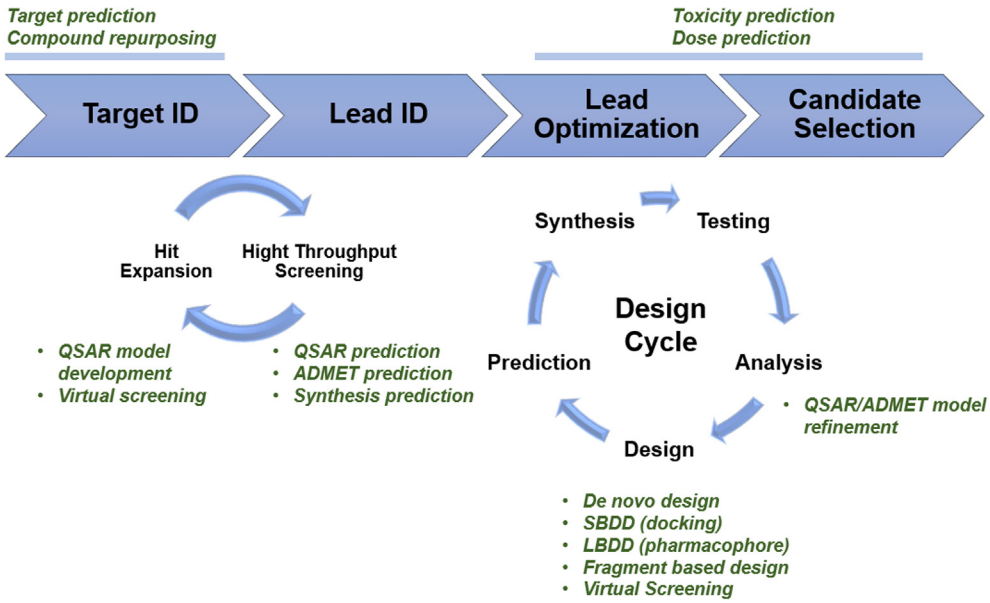

Fig. 2. Application of artificial intelligence and machine learning tools in the early stages of the drug discovery process. SBDD: structure-based drug design, LBDD: ligand-based drug design.