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

From chemoinformatics to deep learning: an open road to drug discovery

Leonardo LG Ferreira^{*,1} & Adriano D Andricopulo^{**,1}

¹Laboratory of Medicinal & Computational Chemistry, Center for Research & Innovation in Biodiversity & Drug Discovery, Physics

Institute of Sao Carlos, University of Sao Paulo, Av. Joao Dagnone 1100, 13563-120 Sao Carlos, SP, Brazil

*Author for correspondence: Tel.: +55 163 373 9874; Fax: +55 163 373 9881; leonardo@ifsc.usp.br

**Author for correspondence: aandrico@ifsc.usp.br

“There has been an exponential rise in the amount of data produced over the last two decades, a phenomenon commonly referred to as big data. In this context, the development of tools to identify patterns and correlations from massive amounts of data has become a key requirement. Pharmaceutical companies have tackled this issue by investing heavily in the implementation of powerful *in silico* methods for data handling, storage and analysis”

First draft submitted: 6 September 2018; Accepted for publication: 11 December 2018; Published online: 19 March 2019

Keywords: artificial intelligence • drug design • machine learning • neural networks

The development of new drugs to address novel and known diseases is one of today's most important drivers for pharmaceutical research. Drug research and development (R&D) is a long and costly process which requires cutting-edge infrastructure and highly qualified human resources. The area of medicinal chemistry and drug design has experienced groundbreaking changes prompted by contemporary scientific and technological innovations. This sector has increasingly absorbed into its R&D settings technologies derived from the breakthrough findings of molecular biology and its related omics disciplines, such as genomics, proteomics, metabolomics and chemogenomics. Apart from promoting key advances in the knowledge of the molecular bases of diseases, these areas have fostered important developments in the technical apparatus used in drug discovery. Among the technologies which have particularly impacted drug R&D are the automation provided by high-throughput screening and the development of organic synthesis methods, such as combinatorial chemistry. Although these improvements have led to a notable growth of innovation in the pharmaceutical industry, they have also increased the complexity of the drug development process.

From chemoinformatics to big data

There has been an exponential rise in the amount of data produced over the last two decades, a phenomenon commonly referred to as big data. In this context, the development of tools to identify patterns and correlations from massive amounts of data has become a key requirement. Pharmaceutical companies have tackled this issue by investing heavily in the implementation of powerful *in silico* methods for data handling, storage and analysis.

The expression ‘artificial intelligence,’ originally coined by John McCarthy in 1956, is used to designate computational systems which are devised to make decisions on a particular task [1]. In the context of chemistry, the more specific term ‘chemoinformatics’ is used to refer to the handling of data related to molecular properties and structure [2]. Among its uses, chemoinformatics is widely applied in the construction of chemometric models for the identification of quantitative structure–activity and structure–property relationships (QSAR and QSPR, respectively). The onset of chemoinformatics can be traced to the mid-1950s when a paper by Ray and Kirsch, published in 1957, described the first substructure searching algorithm [3]. Substantial progress was made in the following decades, however, the inauguration of the big data era in the 2000s demanded a sharp evolution of the existing tools toward increasingly powerful methods.

newlands
press

Machine learning

Machine learning (ML) technologies are among the hottest topics in chemoinformatics. The bases of its modern configuration were developed between the 1960s and the 1980s, and among its several categories, the artificial neural networks (ANNs) have gained major prominence [4]. These algorithms seek to mimic the human brain's connections structure, so that computers can capture data and categorize information in a similar way [5]. To that end, ANNs are trained with a set of events so that they are able to make predictions and make decisions on related occurrences. The learning process is enabled by feedback loops, which allow the system to adapt itself and modify its strategy after being informed whether its decisions were successful or not. This is done by calculating the error between the predicted and actual values using the so-called 'back propagation algorithm', and then updating the weights and biases configured in the network [6].

The evolution of the ANNs has culminated in the development of deep learning (DL) systems, so named to contrast them with the conventional shallow networks. Performing better than the traditional ANNs, these algorithms have successfully been applied in multiple areas, including voice and image recognition, natural language processing, robotics and marketing [7]. DL networks can be divided into several categories, such as convolutional neural networks, recursive neural networks and deep autoencoder systems. Regardless of all the hype generated around these methods, what distinguishes deep nets from the shallow ones? In short, DL has been outlined to solve problems involving massive amounts of data, that is, to deal with the big data problem [8]. To fulfill this task, DL networks rely on a more complex architecture. All ANNs are made of several processing layers, each one dedicated to a different function: an input layer; hidden layers and an output layer. Each layer contains nodes, or neurons, which are the basic computing units. The input layer takes in the information, then a weighted linear combination of the output from its nodes is forwarded to the hidden layers. The hidden layers carry out nonlinear data transformations using an activation function. Finally, the output layer uses its activation function to generate an answer to the proposed problem [9]. DL systems are more complex because they have a higher number of neurons per layer along with multiple hidden layers, whereas shallow ANNs have no more than two layers of this type. Algorithmic improvements, such as the implementation of different transfer functions and regularization techniques, have enabled DL systems to overcome common problems of the conventional ANNs, for instance, the overfitting [8]. Nevertheless, the presence of many hidden layers and neurons per layer requires considerably more computing power and time, a difficulty that can be minimized with the use of graphics processing units. In addition, deep nets are particularly prone to what is called the gradient vanishing problem, which renders model optimization more difficult [9].

Deep learning & drug research & development

The use of DL in drug R&D is a recent phenomenon and its potential has been deeply investigated both in the academic and industrial environments. Research has focused on a variety of areas, including QSAR and QSPR modeling; analyses of synthetic routes; automated molecular design; image recognition and the prediction of mechanisms of action [10–12]. As to the design of novel compounds, which is a laborious process particularly prone to failure, DL has outperformed other ML tools; deep nets have proven valuable to create chemically valid and synthetically accessible molecules with suitable properties for drug discovery purposes [13].

Lusci *et al.* were among the pioneers in using DL for predicting physicochemical properties of small molecules. In a paper published in 2013, they reported on the use of recursive DL to predict aqueous solubility [14]. Aliper *et al.* described in 2016 a DL approach to use transcriptional response data as input to construct a model for the categorization of compounds into therapeutic classes. The DL network showed high accuracy in classifying the dataset compounds, outperforming shallow ANNs that were based on conventional ML methods, such as support vector machine [15]. In a joint effort, Stanford University and Google took part in a survey regarding how screening data from different diseases could be used in virtual screening campaigns. Nearly 40 million data points obtained against more than 200 disease-related targets were used for training a multitask DL network. The results demonstrated predictive accuracies far superior to that produced by conventional single-task systems [16]. As to image recognition, deep networks have been used in high-content imaging systems to analyze cellular alterations upon treatment with screening compounds. Graphics processing units have been particularly useful in this area given their capacity in handling matrices of values and thus recognize digital images. By segmenting, annotating and analyzing sets of images, the system perceives drug-induced phenotypic changes and provides insights into the mechanism of action of compounds [12]. Gomez-Bombarelli *et al.* proposed a method for automated molecular design by applying the autoencoding concept to create novel chemical structures [17]. A DL net was fed with

thousands of structures from the ZINC database to derive a set of coupled functions. By undertaking operations in the latent space, such as vector decoding, perturbation of the known chemical structures and interpolation among structures, the network succeeded in creating novel molecules with drug-like properties. Pursuing the *de novo* design of compounds, Popova *et al.* developed a deep ANN called ReLeaSE to create novel structures which have drug-like features. The method relies on the integration of deep and reinforcement learning approaches and on two DL networks that were trained separately. When combined, these two ANNs can be used to propose new compounds based solely on simplified molecular-input line-entry system strings [18].

Concluding remark & future perspective

Deep learning ANNs still have much to prove in the drug R&D realm. Notwithstanding, the pharmaceutical industry, which has an estimated cost of US\$ 2.6 billion to develop a single drug, and failure rates of 90% between clinical development and approval, is particularly interested in the topic [19]. Some major companies have turned to DL to speed up their R&D processes and increase productivity. In the search for novel drugs for cancer, Pfizer and Roche have associated with IBM. Sanofi and GSK have joined Exscientia to pursue the optimization of automated molecular design methods. AstraZeneca has collaborated with Berg Health to seek for novel therapies for neurological conditions such as Parkinson's disease. Given the recent character of these partnerships and considering that DL is taking its first steps into drug discovery, it is premature to anticipate its impact on the pharmaceutical industry. The drug approvals in the next decade will likely demonstrate if the current DL race will trade-off the investment and bring about reduced attrition rates and more effective medicines.

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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

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

References

- Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 349(6245), 255–260 (2015).
- Pirhadi S, Sunseri J, Koes DR. Open source molecular modeling. *J. Mol. Graph. Model.* 69, 127–143 (2016).
- Willett P. Chemoinformatics: a history. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1(1), 46–56 (2011).
- Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 375(13), 1216–1219 (2016).
- Ding S, Li H, Su C, Yu J, Jin F. Evolutionary artificial neural networks: a review. *Artif. Intell. Rev.* 39(3), 251–260 (2013).
- Denève S, Machens CK. Efficient codes and balanced networks. *Nat. Neurosci.* 19(3), 375–382 (2016).
- Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov. Today* 23(6), 1241–1250 (2018).
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 521(7553), 436–444 (2015).
- Jing Y, Bian Y, Hu Z, Wang L, Xie XS. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS J.* 20(3), 58 (2018).
- Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* 30(8), 595–608 (2016).
- Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3(5), 434–443 (2017).
- Gawehn E, Hiss JA, Brown JB, Schneider G. Advancing drug discovery via GPU-based deep learning. *Expert Opin. Drug Discov.* 13(7), 579–582 (2018).
- Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H. Application of generative autoencoder in *de novo* molecular design. *Mol. Inf.* 37(1–2), 1700123 (2018).
- Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* 53(7), 1563–1575 (2016).
- Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13(7), 2524 (2016).
- Dahl GE, Jaitly N, Salakhutdinov R. Multitask neural networks for QSAR predictions. arXiv[stat.ML] (2014) <http://arxiv.org/abs/1406.1231>

17. Gómez-Bombarelli R, Wei JN, Duvenaud D, *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci.* 4(2), 268–276 (2018).
18. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design. *Sci. Adv.* 4(7), eaap7885 (2018).
19. Fleming N. How artificial intelligence is changing drug discovery. *Nature* 557(7707), S55–S57 (2018).