Mini Review

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2019.23.003831

Applications of Machine Learning in Drug Discovery

Mingbo Zhang[1]*, Huipu Han[1], Zhili Xu[1] and Ming Chu[2]

[1]*College of Pharmacy, PR China*

[2]*Department of Immunology, PR China*

***Corresponding author:** Mingbo Zhang, College of Pharmacy, Liaoning University of Traditional Chinese Medicine, PR China

ARTICLE INFO

Received: November 11, 2019

Published: November 18, 2019

Citation: Mingbo Zhang Huipu Han, Zhili Xu, Ming Chu. Applications of Machine Learning in Drug Discovery. Biomed J Sci & Tech Res 23(1)-2019. BJSTR. MS.ID.003831.

Abstract

Abbreviations: ML: Machine Learning; DL: Deep Learning; AI: Artificial Intelligence; SVM: Support Vector Machine; CADD: Computer Aided Drug Design; RNN: Recursive Neural Networks; LDA: Linear Discriminant Analysis; CNNs: Convolutional Neural Networks;

Mini Review

Recent trend in drug discovery has been marked for the escalating cost and lowering rates of getting approved. In average, it costs upwards of $2.5 billion and about ten years to bring a new drug to the market [1]. Nearly 90% of drug candidates obtained with vast expense will fail somewhere between phase I trials and regulatory approval [2]. So, there is an urgent need for new solution to improve the efficiency of drug discovery process for pharmaceutical industry. In recent years, machine learning (ML) technique has gained a rapid development. Especially, the advent of deep learning (DL) enables the artificial intelligence (AI) to overwhelm human being in certain specific applications such as chess game and image recognition, marked by the victory of Alpha Go over the world strongest human Go player in 2016. Today, ML is widely applied in every aspect of human's social and industrial activity, such as identification of spam email, handwritten word recognition, news recommendation, autonomous driving, medical image analysis, etc. In pharmaceutical industry, ML has become one of the most important and rapidly evolving tools in computer-aided drug discovery, being involved in almost every stage in drug development [3]. There are already several specific and detailed reviews on the applications of ML techniques in drug discovery [3,4]. Here, we present a mini review with special focus on drug target identification and validation, drug design and optimization, and drug toxicity prediction.

Drug Target Identification and Validation

Identification of drug target is an important task in initialing a drug discovery pipeline. Modern biology has accumulated large amounts of human genetic information as well as transcriptomic, proteomic and metabolomic data, which renders it feasible to apply ML to identify drug target. For example, by analyzing the gene expression profile of young and old human skeletal muscle with ML approach, Mamoshina et al. [5] identified a panel of tissue-specific biomarkers of aging, which showed good correlation with the actual age values of muscle tissue samples [5]. Similarly, Jeon et al. built a classifier with support vector machine (SVM) algorithm to identify drug targets for breast, pancreatic and ovarian cancers, utilizing biological information including gene essentiality, mRNA expression, DNA copy number, etc. as classification features. The predicted drug targets were validated by the strong anti-proliferative effects of their inhibitors [6]. Target identification with ML is also useful for diagnosis and treatment of rare diseases, which usually lack effective treatment strategies. IJzendoorn et al. [7] performed machine learning analysis on transcriptome sequencing data, thereby uncovering diagnostic biomarker, prognostic gene and identifying potential novel therapeutic targets for soft tissue sarcomas, a group of rare cancers [7]. In addition to predicting the potential target for specific disease, ML approaches can also be utilized to unravel the common features of drug targets. Using amino acid composition and property group composition as features, Kumari et al. build a model with ensemble classification learning method-rotation forest to distinguish drug target from non-drug target, which proves to be useful for novel drug target identification [8]. In conclusion, machine learning may serve as powerful a tool to speed up target identification and validation.

Drug Design and Optimization

The ultimate goal of drug discovery is to bring new drugs to clinic to treat diseases. Once a target has been identified, the next issue is how to efficiently design and optimize chemical structures that will alter the disease state by modulating the activity of the identified target. In the past decades, computer aided drug design (CADD) has offered valuable tools for identifying active drug candidates, including molecular docking and quantitative structure-activity relationship (QSAR). With the rapid explosion of chemical and biological databases as well as the advance in ML algorithms, ML has become an alternative CADD tool for drug design and optimization [3]. For example, based on random forest (RF) algorithm, a novel score function was proposed to predict protein-ligand binding affinity, which outperformed other 16 classical scoring functions with accuracy increasing with the size of training dataset [9]. ML can also be applied to design inhibitors against non-molecular target. Cruz et al. developed ML models with k-nearest neighbor, RF and SVM algorithms using nuclear magnetic resonance data as features to identify molecules capable of inhibiting growth of cancer cell [10]. The advent of deep learning (DL) methods significantly boost predictive power of ML approaches. For example, in the Merk Kaggle, the DL outperformed RF approach using 2D molecular descriptors for 13 of 15 arrays [11]. Another advantage of DL is that it can be employed to optimize novel chemical structures towards desired properties. Olivecrona et al. designed a model based on recursive neural networks (RNN), which is capable of generating novel compounds with optimized parameters including bioactivity, solubility, pharmacokinetic properties and so on [12].

Prediction of Drug Toxicity

Currently toxicity is the major reason for drug candidate failure during development and clinical trials and is responsible for two-thirds of the drugs pulled off the market [13]. So, it is essential to screen out compounds with the potential toxicity as early as possible to save the capital and labor devoted to the preclinical and clinical investigation [14]. One way to achieve this goal is to develop accurate methods for toxicity prediction. Initially the drug toxicity was predicted with QSAR methods, which build quantitative relationships between chemical structure or properties and drug toxicity [15]. The assumptions of linearity as well as the sensitivity to data dimensionality inherent in the early QSAR models limited their predictability. Currently, massive amount of newly available data makes it a rational choice to turn to ML for the prediction of drug toxicity. Researchers have used a combination of algorithms including k-NN, SVM, RF and DL algorithms to predict toxicity [16]. It was showed that the commonly used ML algorithms such SVM, RF, linear discriminant analysis (LDA) and neural network are unsuitable to process imbalanced Tox Cast data [17]. Fortunately, DL method proved to be a qualified method to treat such imbalanced data. For example, Xu et al. [18] built a live injury (DILI) prediction model with DL based on chemical structure data, which performed better than the DILI models reported previously [18]. In another example, convolutional neural networks (CNNs), a subclass of DL networks has been successfully used to predict toxicity in terms of images of cell pretreated with different drugs [19].

Concluding Remarks

Machine learning has received much attention as a powerful tool for uncovering patterns hidden in data. With the exponential growth of chemical and biological datasets over the past decades, machine learning algorithms such RF, SVM and LDA has been successfully applied to drug discovery process, as described above. Deep learning algorithms showed better performance on property prediction compared to the classic ML algorithms. However, there are still issues that deserves further study. One is the quality of training data, which is a crucial factor for the performance of resulting prediction model. Currently, the public accessible datasets such as Chem BL [20] and Pub Chem [21] are generally built by collecting data from different public literatures. Consequently, the inconsistency in the data collected this way is inevitable, which may ruin the resulting ML model. Here, further study is needed to present systematic, diverse, accurate databases as training dataset for building ML model. The other issue is about the interpretability of ML model. Recent revolution in deep learning networks makes it a promising tool for remarkable predictability. Unfortunately, the DL models are so complicated that their predictions cannot be interpreted or explained in physical or chemical terms, the so-called "black box", which prevents drug designer from gaining insight into the prediction. So, a novel DL algorithm with a balance between predictability and interpretability will be expected in the future.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.

References

1. Di Masi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J Heath Econ 47: 20-33.
2. Wong CH, Siah KW, Lo AW (2018) Estimation of clinical trial success rates and related parameters. Biostatistics 20(2): 273-286
3. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, et al. (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18(6): 463-477.
4. Lavecchia A (2015) Machine-learning approaches in drug discovery: Methods and applications. Drug Discov Today 20(3): 218-331.

5. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, et al. (2018) Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target identification. Front Genet 9: 242.

6. Jeon J, Nim S, Teyra J, Datti A, Wrana JL, et al. (2014) A Systematic Approach to Identify Novel Cancer Drug Targets Using Machine Learning, Inhibitor Design and High-Throughput Screening. Genom Medi 6(7): 57.

7. Van IJzendoorn DGP, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, et al. (2019) Machine Learning Analysis of Gene Expression Data Reveals Novel Diagnostic and Prognostic Biomarkers and Identifies Therapeutic Targets for Soft Tissue Sarcomas. PLoS Comput Biol 15(2): e1006826.

8. Kumari P, Nath A, Chaube R (2015) Identification of Human Drug Targets Using Machine-Learning Algorithms. Comput Biol Med 56: 175-181.

9. Ballester PJ, Mitchell JBO (2010) A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. Bioinformatics 26(9): 1169-1175.

10. Cruz S, Gomes S, Borralho P, Rodrigues C, Gaudêncio S, et al. (2018) In Silico HCT116 Human Colon Cancer Cell-Based Models En Route to the Discovery of Lead-Like Anticancer Drugs. Biomolecules 8(3): 56.

11. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V (2015) Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. J Chem Inf Model 55(2): 263-274.

12. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular De-Novo Design through Deep Reinforcement Learning. J Cheminform 9(1): 48.

13. Onakpoya IJ, Heneghan CJ, Aronson JK (2016) Worldwide Withdrawal of Medicinal Products Because of Adverse Drug Reactions: A Systematic Review and Analysis. Crit Rev Toxicol 46(6): 477-489.

14. Li AP (2004) Accurate Prediction of Human Drug Toxicity: A Major Challenge in Drug Development. Chem-Biol Interact 150(1): 3-7.

15. Patlewicz G, Fitzpatrick JM (2016) Current and Future Perspectives on the Development, Evaluation, and Application of in Silico Approaches for Predicting Toxicity. Chem Res Toxicol 29(4): 438-451.

16. Basile AO, Yahi A, Tatonetti NP (2019) Artificial Intelligence for Drug Toxicity and Safety. Trends in Pharmacol Sci 40(9): 624-635.

17. Grenet I, Merlo M, Comet J, Tertiaux R, Rouquié D, et al. (2019) Stacked Generalization with Applicability Domain Outperforms Simple QSAR on in Vitro Toxicological Data. J Chem Inf Model 59(4): 1486-1496.

18. Xu Y, Dai Z, Chen F, Gao S, Pei J (2015) Deep Learning for Drug-Induced Liver Injury. J Chem Inf Model 55(10): 2085-2093.

19. Jimenez Carretero D, Abrishhami V, Fernández de Manuel L, Palacios I, Quílez Álvarez A, et al. (2018) Tox_(R) CNN: Deep LearningBased Nuclei Profiling Tool for Drug Toxicity Screening. PLoS Comput Biol 14(11): e1006238.

20. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, et al. (2014) The ChEMBL Bioactivity Database: An update. Nucleic Acids Res 42: D1083-D1090.

21. Wang Y, Suzek T, Zhang J, Wang J, He S, et al. (2014) PubChem Bioassay: 2014 Update. Nucleic Acids Res 42: D1075-D1082.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2019.23.003831

Mingbo Zhang. Biomed J Sci & Tech Res