International Conference on Intelligent Technologies & Science - 2022
(ICITS-2022)

A Brief Review on Machine Learning Approaches in Drug Discovery

Arvind Kumar[1], Sushil Kumar Maurya[2]
Computer science and Engineering Department,
Bansal Institute of Engineering and Technology, Lucknow
Arvindrajbhar323@gmail.com

Abstract: This review provides the feasible literature on drug discovery through ML tools and techniques that are enforced in every phase of drug development to accelerate the research process and deduce the risk and expenditure in clinical trials. Machine learning techniques improve the decision-making in pharmaceutical data across various applications like QSAR analysis, hit discoveries, de novo drug architectures to retrieve accurate outcomes. Target validation, prognostic biomarkers, digital pathology are considered under problem statements in this review. ML challenges must be applicable for the main cause of inadequacy in interpretability outcomes that may restrict the applications in drug discovery.

Keywords: Artificial intelligence, Drug discovery, Machine learning, Target validation.

1. Introduction:

Drug discovery is a very complicated process as it involves a huge investment of time and money. On average, it takes 6 to 12 years with an investment of 500 million to 1 billion (in US dollars) to identify a drug for fighting against a target. However, even after a huge struggle, the success rate is very low. Many long-term research projects may end up fruitless resulting in wastage of enormous efforts. Blockbuster drugs are the drugs that are prescribed for the common medical problems like cold, diabetes, high blood pressure, asthma and u. They are extremely profitable in the pharmaceutical industry. They bring revenues greater than 1 billion per year and a profit of more than 1 million a day (in dollars). However, it can also result in problems for the company if the drug shows any side effects. Usually, the patents on drugs expire resulting in competition from less expensive equivalents. The process of drug discovery is therefore highly complicated and risky activity but is always motivated by the benefits it could do to millions of people suffering from various diseases. The detailed process of drug discovery, illustrated, in Figure 1.1 is as follows:

1. The first stage is the target discovery. In this stage, we decide on the target on which the drug in development should act upon to suppress the growth of the disease. The target gives us the deeper understanding of the genes, proteins, Ribonucleic Acid (RNA) or anything that get effected when attacked by a parasite.
2. The second stage is the target validation phase. In this phase, the discovered target is validated to make sure that the drug in development deals with the correct target.
3. The third stage is the lead discovery. This phase involves in synthesizing and isolating the designed chemical compounds meant to interact with the specified targets. This stage involves the use of chemistry, assay development for screening the selected compounds.

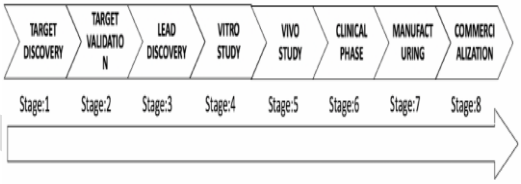

Figure 1. Process of Drug Discovery

2. Related Work:

Aging studies are made on organisms such as Caenorhabditis elegans (C. elegans) worm, yeast, mice and (Buffenstein et al., 2008). Among them, C. elegans is highly used for conducting assessments on anti-aging and drug intervention due to its short lifespan, stereotypical development and small size (Kaletta and Hengartner, 2006) (Lucanic et al., 2013) (Carretero et al., 2017). There has been a decent increase in the number of works done on the identification of compounds that increase the lifespan. Latest research enabled us to develop compounds that enhance the longevity of caloric restriction (Fontana et al., 2010). In (Calvert et al., 2016), a pharmacological network is constructed by Ye et al. with the help of a connectivity map to identify drugs with overlapping gene expression profiles.

It helped to identify 60 compounds that improve longevity for C. elegans (Ye et al., 2014). Ranking of drug like compounds to modulate aging in C. elegans has been proposed in (Ziehm et al., 2017). This ranking method is based on genetic information, information on proteins associated with aging in an organism, 3D protein structure, homo-logy and sequence conversation between them and compound activity information. In (Ding et al., 2017) and (Carretero et al., 2015), a review on anti-aging and pharmacological compound classification on C. elegans lifespan has been discussed. The process of drug discovery includes several steps from selecting chemical compound candidates to clearing all drug requirement tests. It is also time consuming, expensive and labor-intensive. Recent advancements in machine learning is helping out in prediction of the chemical properties in drug discovery community. In (Zernov et al., 2003), Support Vector Machine is used for predicting the drug-likeness and agrochemical-likeness for large number of compounds. In

(Putin et al., 2016), human chronological age was predicted using Deep Neural Network. In (Fabris et al., 2018), prediction of aging related genes was done using random forests.

Different possibilities of involvement of machine learning and artificial intelligence in longevity and anti-aging discoveries was discussed in (Batin et al., 2017). Machine learning was not greatly used for compound prediction on longevity. To the best of our knowledge, only a few studies tried to identify chemical compounds that effects longevity (Putin et al., 2016)(Barardo et al., 2017) (Fabris et al., 2017). Barardo et al. worked on classification of chemical compounds into two classes which are the ones that effect longevity and the ones that do not, using random forests (Barardo et al., 2017). They used random forest feature importance as a parameter to select the most important features among the two types of features which are chemical descriptors, gene ontology terms. Then, they implemented random forests for classification. Their results showed that impact of chemical descriptors was high compared to GO. However, using both chemical descriptors and Gene Ontology slightly improves the classification accuracy.

Information provided by features in a data set play a major role in classifying each instance into different classes. In many cases, a compromise on relevant and redundant features is observed in the data. Redundant features slow down the process of learning resulting in a decrease in accuracy. Feature selection helps us to identify and eliminate these redundant features helping us to improve the performance of the classification (Dash and Liu, 1997) (Xue et al., 2013). Exponential increase in the size of search space with the size of features makes feature selection much more challenging. Due to this, search methods such as greedy search or heuristic-based search have been proposed (Mao and Tsang, 2013). But these methods are prone to suffer from the local optima problem. Feature selection can be done using two approaches: 1) methods that are independent of a learning algorithm and do not consider classifier performance (Moradi and Rostami, 2015), and 2) wrapper methods which learns an algorithm in the feature selection process (Gasca et al., 2006) (Wang et al., 2008). Wrapper methods can enable us to achieve higher predictive accuracy (Dash and Liu, 1997).

Determining Drug-Target Interaction (DTI) is a major part in the area pharmaceutical sciences (Fakhraei et al., 2014). The process of drug discovery involves costs around $1.8 billion with a duration which may extend beyond 10 years (Ciociola et al., 2014). Thus, the drug-target interactions prediction helps in narrowing down the search area helping out the biologists. The main step in the process of drug discovery is to recognize the targets related to the drugs. These targets are mostly the proteins that can be drugged, and the ones related to diseases. The prediction of drug-target interactions aims at identifying the possible new targets for the existing drugs. It basically guides us through a proper experimentation process. There are many types of protein targets which include enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors. These classes can modulate their function by interacting with various ligands. Thus, analyzing the genomic space produced by these classes of proteins, helps us to accurately estimate the possibility of an interaction. DTI can be used either for drug discovery or for re-positioning which is reusing available drugs for new targets. Approaches to predicting DTI can be classified into one of three categories: 1) Ligand-based, 2) Docking-based and 3) Chemogenomic approach. Prediction of DTI based on target proteins' ligand similarity is the Ligand-based approach (Keiser et al., 2007) (Keiser et al., 2009). 3D structure information of a target protein can help in estimating the likelihood of its possible interaction with certain drug. This is based on their binding capacity and strength (Cheng et al., 2007) (Morris et al., 2009). This method is the Docking-based approach. Lastly, if the chemical information of the drugs, genomic information from proteins and already identified DTIs are used, then this is the chemogenomic approach (Mousavian and Masoudi-Nejad, 2014) (Ding et al., 2013). A target with small number of binding ligands often leads to poor DTI predictions in a ligand-based method. This is a drawback in this method. Similarly, docking-based method is based on availability of 3D structures for target proteins and is also time consuming. These disadvantages made the chemogenomic approach more popular for identification of DTI in the recent times. This approach models the DTI problem as a machine learning problem and often builds a classifier which is trained by an available interaction data. This classifier is used to predict the unknown interactions (Ding et al., 2013).

Different techniques are employed in chemogenomic approach. Some of these include bipartite graph (Bleakley and Yamanishi, 2009)(Lu et al., 2017), recommendation systems (Alaimo et al., 2016) and supervised classification problem (Wen et al., 2017). But considering the data, we can see that there will be only a few positive interactions and the remaining possible interactions are unknown. For example, of the 35 million drug compound possible candidates, total number of positive drug interactions could just be 7000 (Bolton et al., 2008). Computational chemogenomic approaches can be classified into feature-based and similaritybased methods. Feature-based methods have features as inputs for a set of instances defined by a particular class label. The instances are generally the drugs and the features are the targets. The class label is a binary value indicating the presence of a possible interaction. Some of the feature-based methods like decision trees, random forests (Breiman, 2001) and support vector machines (Cristianini and Shawe-Taylor, 2000) are used for classification purposes. Generally, Random Forest and Support Vector Machine are used for the classification of drug-target interactions (Yu et al., 2012).

Similarity-based methods take the similarity matrices of drugs and targets as inputs including the interaction matrix which indicates the drug-target pairs that most likely interact with each other. These approaches assume the common features among the drugs and targets can determine the presence of an interaction between a drug and target. The main focus is on the structural similarity between the drugs and targets leaving

out the remaining unknown attributes. This gives an opportunity for easy implementation of the similarity indices on the networks without any prior information. To implement the similarity-based algorithms, many similarity indices were used by (Lu et al., 2017). There are two ways to classify the similarity indices as local similarity indices and global similarity indices. The local similarity indices are node-dependent i.e. they require information related to neighbourhood of the network. On the other hand, global similarity indices are dependent on the path through the entire network topology (Lu et al., 2017). Use of the similarity indices outperform many random link predictors. A few examples of the similarity indices include Common Neighbours (CN), Jaccard Index, Preferential Attachment (PA) and Katz Index (Lu et al., 2017).

Drug-target link prediction is of high interest in the recent times. The application of machine learning in this challenging field of study makes it more promising and researchers have been trying to explore in depth in this field. (You et al., 2018) developed a lasso-based regularized linear classification method to predict the DTI overcoming the high-dimensional nature of the drug target data with huge number of features and small number of samples. A recent use of stacked auto encoder from the drug molecular structure and protein sequence to extract sufficient information was given by (Wang et al., 2018). DINES, a web server defined as drugtarget interaction network inference engine based on supervised analysis is used to predict unknown DTI for different biological data. It predicts with the help of machine learning methods integrating with the heterogeneous biological data in compatibility with the KEGG data (Yamanishi et al., 2014).

A framework was proposed by (Fakhraei et al., 2014) to work with a bipartite graph of drug-target interactions along with similarity measures. This used probabilistic soft logic to make predictions based on triad and tetrad structures. Enhancing the similarity measures to involve the non-structural information and to handle the possible missing interactions was done by (Shi et al., 2015). Matrix-factorization has been used in many research to identify new drug-target interactions (G□onen, 2012) (Eslami Manoochehri and Nourani, 2018) (Zheng et al., 2013). (Liu et al., 2016) used matrix-factorization method to focus on determining the probability of a drug-target interaction. In this method, the properties of drugs and targets were represented using their specific latent vectors. A modification to this method which uses Bayesian optimization was developed by (Ban et al., 2017). It uses the neighborhood regularized logistic matrix factorization for DTI prediction. A multiple kernel link prediction on bipartite graphs was proposed by (Nascimento et al., 2016). In this method, relevant kernels are selected based on the weights which define the importance of a drug-target link. Bipartite Local Model (BLM) was extended to DTI prediction with hubness-aware regression in (Buza and Peska, 2017). It builds a projection-based ensemble of BLMs using similarity space of drugs and targets. In (Lan et al., 2016), the unknown links are treated as unlabeled samples. A majority voting method is used to decide on the label of these unlabeled samples using the method of Positive-unlabeled learning for drugtarget prediction (PUDT). This unlabeled data is divided into reliable negative samples and likely negative samples with the help of their similarity information. Weighted SVM is used to then classify this data. We can also have positive unlabeled data similar to negative unlabeled data. To deal with this data, two group of approaches were proposed. One identifies the reliable negatives among the data and use the positives and these negatives to build the model. The other group directly predicts the positive unlabeled data (Liu et al., 2017).

Typically, there is a lot of unbalanced data in the drug-target interaction database. Weighted profile can be used to determine the drug profiles with defines weights as similarities between drug-drug (Yamanishi et al., 2008). The nearest profile is an approach that predicts the interaction profile of new drugs. (van Laarhoven and Marchiori, 2013) developed a simple weighted nearest neighbor procedure for predicting the interacting pairs with high accuracy. A network based inference helps in building a predictive model similar to a network where the nodes are represented by drugs and targets. Interacting drugs and targets are represented as edges (Cheng et al., 2012).

Link prediction has found a noticeable place in many applications as social networking, e-commerce, etc. (Wang et al., 2015). Heuristic approaches have been mostly used for the identification of any possible links. In the field of medicine, to predict a possible DTI, (van Laarhoven and Marchiori, 2013) has used common weighted nearest neighbors method. Some methods used shortest path analysis between the drugs and targets. These heuristic methods generally, do not reveal much information about the in-depth relations between drugs and targets. In this work, a bipartite graph was first constructed with the available data. Then we worked on proper sampling of the huge number of negative samples in the data sets. We have used Weisfeiler-Lehman Neural Machine (WLNM) algorithm for creating the adjacency matrix which represents the interactions between potential drugs and targets. This adjacency matrix is fed to a machine learning model. We used 10-fold cross validation and AUC to estimate the efficiency. We compared the performance between Neural Network, Support Vector Machine and logistic regression models.

The crucial step of drug development in silico is consisting of multiple biological sources for predicting drug–protein interactions. Here complications can be seen in large predictions, which relied on the countless unknown interactions. Therefore, semi-supervised training techniques should be used to address these unlabelled and labeled date complications. Usually, only labeled data will produce better results. In addition, the semi-supervised technology integrates chemical structure, drug–protein interaction network data, and genome sequence data. Finally, in this article, drug–protein interactions of various data sets such as ions, enzymes, and nuclear receptors provided well predictable results (Xia et al. 2010).

Drugs have an important priority in therapeutic activity, which is regulated by protein interactions. The drug–protein interaction database (DPI) focuses primarily on therapeutic

protein targets, while knowledge of non-targets has been limited and resolved. Thus, computational techniques can fill the knowledge gap for predicting protein targets for distributed drug molecules. In that study, the pool of 35 predictors had a major impact on the similarity between protein and drug targets. Drug structure, target sequence, and drug profile are three types of similarity developed from the results of 35 predictors. Finally, the significant content, relationships, and implications between database sources are of great importance for therapeutic activity (Wang and Kurgan 2020). In drug repurposing, the unexpected detection of drug–protein interactions is essential. Thus, the dominant drug may be useful for repurposing, while drug side effects are unavoidable and about 1,000 human proteins can cause critical side effects. The proteomic scale method was used to predict side effects and protein goals. FINDSITEcomb is used to predict drug–protein interactions. The estimates showed greater disruption with a mean of 329 human targets for each drug (Zhou et al. 2015).

Usually, a drug effect assessment looks at its biochemical activity. The effectiveness of the therapeutic activity has posed a challenge to be properly coupled with the biochemical activity. The collection of a large amount of data on the effects of cellular drugs was undertaken to fill a gap that has been explored in the extensive content of cellular estimations and while this estimation is classified as a psychotropic drug. Here, the microarray data can be analyzed by applying random trees to the forest and classifying them, providing a profile for the efficiency of biomarker gene expression. Accuracy of 88.9% of the classification tree and 83.3% of the random forest model used this efficacy profile for a drug treatment analysis. Therefore, at the cellular assessment level, general genomic data are acceptable to reconcile the effects of new physiological drugs with clinical applications. Finally, in vitro signatures of gene expression data can identify the effectiveness of therapeutic activities that can help validate targeting and drug development (Gunther et al. 2003).

In drug development, increasing profitability by validating new drugs requires predicting effectiveness and identifying targets. The proximity of medical illnesses helps to reduce the effectiveness of the treatment and also releases drugs that are effective in therapeutic activity. The study treated 78 diseases with 238 drugs to demonstrate the drug's effectiveness in therapeutic activity, as well as problems with gene efficacy and various disorders.

Here the network-based system is used to develop a drug-disease proximate measure that assesses the interactions between the disease and the drug target. Therefore, the proximity of network-based systems makes it possible to predict associations for novel drug diseases, offering a wide range of possibilities for conflict detection and drug repurposing (Guney et al. 2016).

In drug development, the use of biomarkers supports the provision of safety measures that critically determine the biological and analytical indicators of a particular biomarker. In this way, stakeholders can assess and manage whether claims are defended for a particular purpose and whether the desired standards are being met. For shareholders in the implementation of evaluating the experiment agreement, a stakeholder evaluation process is needed to adjust the unique characteristics of the biomarkers, as well as to determine how these innovations are analyzed, integrated, and interpreted, and how improved biomarkers and conventional comparators are measured (Sistare et al. 2010).

In the survey, we found that modern medicines are no safer than older drugs, even though with longer medical trial programs. These trails are placed on the market and impractical inspections are carried out which are not sufficient to be carried out systematically to ensure safety. Previous drug-related signals can help in improving drug safety as well as identify underlying biomarkers, making them more toxic. However, the safety markers can be different for different target systems. However, no other approach can provide assurance that medicines are very safe, but we can develop a common understanding of benefit and risk assessment by communicating with the public (Rolan et al. 2007).

Various deep learning techniques are carried out here to predict the PPI interface and show fantastic results when contrasted with the SVM technique (Du et al. 2016). Thus, the PPI's became more complex to utilize in biological techniques (Falchi et al. 2014; Scott et al. 2016). Each PPI can be a mixture of various residues (Cukuroglu et al. 2014). New PPI can act as a modern class for pharmaceutical targets where disparate for different targets i.e., ion channels, GPCRs (G-Protein coupled receptors), kinases (Higueruelo et al. 2013; Santos et al. 2017). iFitDock is a docking tool used for investigating a few hotspots in PPIs. Further, AI techniques have been utilized for distinguishing structures and hotspots in PPI interface.

AI innovation has a high priority in drug design through the enhancement of ML approaches and the collection of pharmacological data. AI does not rely upon any hypothetical improvements, but it has more essence in transforming medical information into studies like reusable methods. In general, there are different approaches such as Random Forest, Naive Bayesian Classification (NBC), Multiple Linear Regression (MLR), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Probabilistic Neural Networks (PNN), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), etc are considered in the context of ML (Lavecchia and Di Giovanni 2013). In order to gain capability in feature extraction and feature generalization, AI advancements are specifically used as a deep learning technique towards drug design. Also, Fig. 1 shows respective applications which illustrate an outline of AI procedures utilized to respond to drug discovery queries in the review. A scope of classifier and regression strategies i.e. supervised learning techniques utilized to respond addresses desire expectations in continuous or categorical data factors, also unsupervised methods utilized in creating a model which empowers the clustering data.

Many designed features in traditional ML models are performed manually, but deep learning approaches will accelerate various features through available initialized data

automatically because multi-layer feature extraction techniques are used to convert straightforward features into complex features. One advantage of using deep learning approaches was, presence of low quantity generalization blunders, so it recovers more exact results. CNN, RNN, Auto Encoder, DNN, and RBN are considered as different deep learning techniques. Summary of deep learning algorithms can be identified (LeCun et al. 2015; Angermueller et al. 2016; Schmidhuber 2015) and provides detailed information about deep learning techniques which are available in Deep Learning literature (Goodfellow et al. 2016).

In drug discovery and development, many AI calculations are associated to analyse and predict the data. Here, few popular models like SVM, RF, and MLP discuss their effective use in drug discovery.

The review of drug discovery is further categorized on the basis of task performing of ML and their applications like target identification, hit discovery, hit to lead, lead optimization techniques are discussed out. The drug design techniques rely on the databases which are inturn developed based on the different ML algorithms. The precise training, validation, and application of ML algorithms in the drug discovery era provide an enthusiastic outcome by easing the complicated error-prone protocols. The ML techniques are introduced in most of the drug design processes to reduce the time as well as manual interference. The best example is QSAR, in which the huge data collection and training of datasets are considered as rate-limiting steps in defining the ligand-based virtual screening protocols and are now replaced by Denovo design techniques.

Information about the primary amino acid sequences of proteins/enzymes/receptors, both dissolved / insoluble, is stored on the UNIPROT server along with their targets and cellular functions. Based on medicinal chemistry or pharmacological or biochemical studies, the main role of proteins is identified, and this information is also the basic unit for developing the protein folding prediction studies by software or experimental studies. Whereas, the protein folding predictions in the provided aminoacid (UNIPROT) sequence were compared with its experimentally derived PDB homologues which became a hopeful technique to refine the new protein models computationally and is also termed as ”homology modeling”. The homology modeling or comparative modeling is analyzed by the several algorithms which need to be implemented in either software modules (PRIME) or web servers (EXPASY, SWISS-MODEL) will definitely make a decision to predict the secondary structure folding with high accuracy within provided templates. However, the fine-tuning for the obtained homology models or template-based models are again scrutinized by Ramachandran analysis which can be sorted out by commercial modules (PRIME) or web servers (QMEAN, PROCHEK).

3. Conclusion:

The AI technology is utilized in pharmaceutical industries including ML algorithms and deep learning techniques in daily life. ML techniques in drug development regions and health service centers have encountered numerous conflicts, especially in image analysis and omics data. In medical science, ML models predict the trained data in a known framework i.e., the compound structure can perform alternative tools like PPT inhibitors, macrocycles with traditional algorithms. Additionally, deep learning models can be considered the chemical structures and QSAR models from pharmaceutical data which was pertinent for molecules with appropriate properties, because to the forward success rate in clinical trials.

References:

[1] Collins FS, Varmus H. A new initiative on precision medicine. New England J Med 2015;372(9):793–5.

[2] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.

[3] Romond EH, Perez EA, Bryant J, Suman VJ, Geyer Jr CE, Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, et al. Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. N Engl J Med 2005;353(16):1673–84.

[4] Blanco JL, Porto-Pazos AB, Pazos A, Fernandez-Lozano C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. Sci Rep 2018;8(1):1–11.

[5] Munteanu CR, Fernández-Blanco E, Seoane JA, Izquierdo-Novo P, Angel Rodriguez-Fernandez J, Maria Prieto-Gonzalez J, Rabunal JR, Pazos A. Drug discovery and design for complex diseases through qsar computational methods. Current Pharmaceutical Des 2010;16(24):2640–55.

[6] García I, Munteanu CR, Fall Y, Gómez G, Uriarte E, González-Díaz H. Qsar and complex network study of the chiral hmgr inhibitor structural diversity. Bioorganic Med Chem 2009;17(1):165–75.

[7] Liu Y, Tang S, Fernandez-Lozano C, Munteanu CR, Pazos A, Yu Y-Z, Tan Z, González-Díaz H. Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. Expert Syst Appl 2017;72:306–16.

[8] Riera-Fernández P, Munteanu CR, Dorado J, Martin-Romalde R, Duardo- Sanchez A, Gonzalez-Diaz H. From chemical graphs in computer-aided drug design to general markov-galvez indices of drug-target, proteome, drugparasitic disease, technological, and social-legal networks. Current Computeraided Drug Des 2011;7(4):315–37.

[9] Shirvani P, Fassihi A. Molecular modelling study on pyrrolo [2, 3-b] pyridine derivatives as c-met kinase inhibitors, a combined approach using molecular docking, 3d-qsar modelling and molecular dynamics simulation. Mol Simul 2020:1–16.

[10] B. Suay-Garcia, J.I. Bueso-Bordils, A. Falcó, M.T. Pérez-Gracia, G. Antón-Fos, P. Alemán-López, Quantitative structure–activity relationship methods in the discovery and development of antibacterials, Wiley Interdisciplinary

Reviews: Computational Molecular Science e1472.
[11] Fernandez-Lozano C, Gestal M, Munteanu CR, Dorado J, Pazos A. A methodology for the design of experiments in computational intelligence with multiple regression models. PeerJ 2016;4:e2721.
[12] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic acids research 46 (D1) (2018) D1074– D1082..
[13] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47(D1):D1102–9.
[14] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, et al. Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1): D1100–7.
[15] Sterling T, Irwin JJ. Zinc 15–ligand discovery for everyone. J Chem Inform Modeling 2015;55(11):2324–37.
[16] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.
[17] Gramatica P, Sangion A. A historical excursus on the statistical validation parameters for qsar models: a clarification concerning metrics and terminology. J Chem Inform Modeling 2016;56(6):1127–31.
[18] Stephenson N, Shane E, Chase J, Rowland J, Ries D, Justice N, Zhang J, Chan L, Cao R. Survey of machine learning techniques in drug discovery. Current Drug Metabolism 2019;20(3):185–93.
[19] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discovery 2019;18 (6):463–77.
[20] Gramatica P. Principles of qsar modeling: comments and suggestions from personal experience. Int J Quantitative Structure-Property Relationships (IJQSPR) 2020;5(3):61–97.
[21] Cronin MT, Schultz TW. Pitfalls in qsar. J Mol Struct (Thoechem) 2003;622(1– 2):39–51.
[22] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110(18):5959–67.
[23] Valdés-Martiní JR, Marrero-Ponce Y, García-Jacas CR, Martinez-Mayorga K, Barigye SJ, d'Almeida YSV, Pérez-Giménez F, Morell CA, et al. Qubils-mas, open source multi-platform software for atom-and bond-based topological (2d) and chiral (2.5 d) algebraic molecular descriptors computations. J Cheminformatics 2017;9(1):1–26.
[24] Fourches D, Ash J. 4d-quantitative structure–activity relationship modeling: making a comeback. Expert Opinion Drug Discovery 2019;14 (12):1227–35.
[25] Wang T, Yuan X-S, Wu M-B, Lin J-P, Yang L-R. The advancement of multidimensional qsar for novel drug discovery-where are we headed? Expert Opinion Drug Discovery 2017;12(8):769–84.
[26] Lee M, Kim H, Joe H, Kim H-G. Multi-channel pinn: investigating scalable and transferable neural networks for drug discovery. J Cheminformatics 2019;11 (1):46.