



Review

Machine Learning Methods in Drug Discovery

Lauv Patel ^{1,†}, Tripti Shukla ^{1,†}, Xiuzhen Huang ², David W. Ussery ³  and Shanzhi Wang ^{1,*} 

¹ Chemistry Department, University of Arkansas at Little Rock, Little Rock, AR 72204, USA; lhpatel@ualr.edu (L.P.); tshukla@ualr.edu (T.S.)

² Department of Computer Science, Arkansas State University, Jonesboro, AR 72467, USA; xhuang@astate.edu

³ Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA; DWUssery@uams.edu

* Correspondence: sxwang2@ualr.edu

† These authors contributed equally to this work.

Academic Editor: Giosuè Costa

Received: 15 October 2020; Accepted: 9 November 2020; Published: 12 November 2020



Abstract: The advancements of information technology and related processing techniques have created a fertile base for progress in many scientific fields and industries. In the fields of drug discovery and development, machine learning techniques have been used for the development of novel drug candidates. The methods for designing drug targets and novel drug discovery now routinely combine machine learning and deep learning algorithms to enhance the efficiency, efficacy, and quality of developed outputs. The generation and incorporation of big data, through technologies such as high-throughput screening and high through-put computational analysis of databases used for both lead and target discovery, has increased the reliability of the machine learning and deep learning incorporated techniques. The use of these virtual screening and encompassing online information has also been highlighted in developing lead synthesis pathways. In this review, machine learning and deep learning algorithms utilized in drug discovery and associated techniques will be discussed. The applications that produce promising results and methods will be reviewed.

Keywords: machine learning; drug discovery; deep learning; in silico screening

1. Introduction

Advancements in computational science have accelerated drug discovery and development. Artificial intelligence (AI) is widely used in both industry and academia. Machine learning (ML), an essential component in AI, has been integrated into many fields, such as data generation and analytics. The basis of algorithm-based techniques, such as ML, requires a heavy mathematical and computational theory. ML models have been used in many promising technologies, such as deep learning (DL) assisted self-driving cars, advanced speech recognition, and support vector machine-based smarter search engines [1–4]. The advent of these computer-assisted computational techniques, first explored in the 1950s, has already been used in drug discovery, bioinformatics, cheminformatics, etc.

Drug discovery has been based on a traditional approach that focuses on holistic treatment. In the last century, the world's medical communities started to use an allopathic approach to treatment and recovery. This change led to the success of fighting diseases, but high drug costs ensued, becoming a healthcare burden. While quite diverse and specific to candidates, the cost of drug discovery and development has consistently and dramatically increased [5]. As illustrated in Figure 1, the generalized components of early drug discovery include target identification and characterization, lead discovery, and lead optimization. Many computer-based approaches have been used for the discovery and optimization of lead compounds, including molecular docking [6,7], pharmacophore modeling [8], decision forests [9], and comparative molecular field analysis [10]. ML and DL have become attractive

approaches to drug discovery. The applications of ML and DL algorithms in drug discovery are not limited to a specific step, but for the whole process. In this article, we review the ML and DL algorithms that have been widely used in drug discovery.

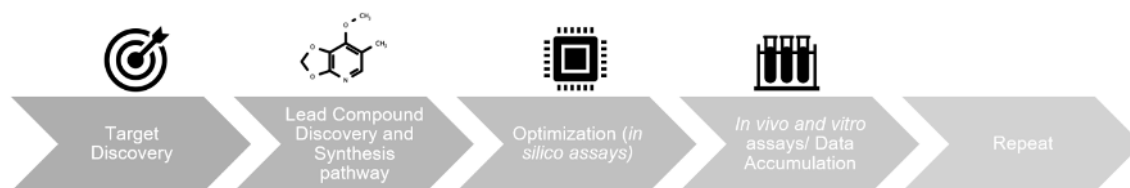


Figure 1. The general steps in drug discovery. Machine learning (ML) and deep learning (DL) algorithms may participate in each of the four steps listed, e.g., by mining proteomic in target discovery, discovering small molecules as candidates in lead discovery, developing quantitative structure-activity relationship models to optimize lead structures for improved bioactivity, and analyzing massive assay results.

2. ML Algorithms Used in Drug Discovery

ML algorithms have significantly advanced drug discovery. Pharmaceutical companies have greatly benefited from the utilization of various ML algorithms in drug discovery. ML algorithms have been used to develop various models for predicting chemical, biological, and physical characteristics of compounds in drug discovery [11–19]. ML algorithms can be incorporated in all steps of the process of drug discovery. For example, ML algorithms have been used to find a new use of drugs, predict drug-protein interactions, discover drug efficacy, ensure safety biomarkers, and optimize the bioactivity of molecules [20–24]. ML algorithms that have been widely used in drug discovery, which include: Random Forest (RF), Naive Bayesian (NB), and support vector machine (SVM) as well as other methods [25–27].

ML algorithms and techniques are not a monolithic, homogeneous subset of AI. There are two main types of ML algorithms: Supervised and unsupervised learning. Supervised learning learns from training samples with known labels to determine labels of new samples. Unsupervised learning recognizes patterns in a set of samples, usually without labels for the samples. The data are usually transformed into a lower dimension to recognize patterns in high-dimensional data using unsupervised learning algorithms prior to recognizing patterns. Dimension reduction is useful, not only, because unsupervised learning is more efficient in a low dimension space but also because the recognized pattern can be more easily interpreted. Supervised and unsupervised learning can be combined as semi-supervised and reinforcement learning, where both functions can be utilized for various data sets [28]. Expansive volumes of data are critical for the development, evolution, and viability of successful ML algorithms in every step of the drug discovery process. The reliance on big high-quality data and known, well-defined training sets are even more essential in precision medicine and therapies within drug discovery. Precision medicine requires a comprehensive characterization of all related pan-omic data: Genomic, transcriptomic, proteomic, etc., to assist in developing genuinely effective personalized medicines. The widespread use of high-throughput screening and sequencing, online multi-omic databases, and ML algorithms, in the past two decades, have created a flourishing environment for many aspects of data generation, collection, and maintenance required for drug development. The advancements of data analytics have successfully attempted to describe and interpret the generated data. This endeavor, supported with ML techniques and integrated databases through multiple software/web-tools (Tables 1–3), is now regularly used for all steps in drug discovery. The ability of new data analytics to synergize with classical approaches and prior hypotheses to produce novel hypotheses and models has proven itself to be useful in applications of repositioning, target discovery, small molecule discovery, synthesis, etc. [29–31]. The information generated within the medical and multi-omic fields is multidimensional. The data is often noisy and heterogeneous in character and source. Using ML methods, like generalized linear models through NB, the issues of analysis and interpretation of multidimensional data may be unburdened. Other ML techniques and

models commonly used in these areas of analysis include regression, clustering, regularization, neural networks (NNs), decision trees, dimensionality reduction, ensemble methods, rule-based methods, and instance-based methods [31,32].

Table 1. Databases used for target discovery.

Databases	Specific Information	Ref.
BRENDA http://www.brenda-enzymes.org	Enzyme and enzyme-ligand information source.	[33]
KEGG http://www.genome.jp/kegg	Database containing genomic information for functional interpretation and practical application.	[33]
PubChem https://pubchem.ncbi.nlm.nih.gov	Database for encompassing information on chemicals and biological activities.	[33]
TTD http://bidd.nus.edu.sg/group/ttd/ttd.asp	Therapeutic Target Database containing encompassing information about the drug resistance mutations, gene expressions, and target combinations data.	[33]
DrugBank http://www.drugbank.ca	Detailed drug data and drug-target information database.	[33]
SuperTarget http://bioinfapache.charite.de/supertarget	Drug-related information databases with more than >300,000 compound-target protein relations.	[33]
TDR targets http://tdrtargets.org	Database containing chemogenomic information for neglected tropical diseases.	[33]
STITCH http://stitch-beta.embl.de	Chemical-Protein interaction networks.	[28]
SMD http://genome-www5.stanford.edu	Database of raw microarray datasets.	[34]
Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo	Database of raw microarray datasets.	[34]
caArray http://array.nci.nih.gov/caarray	Database of cancer-related microarray datasets.	[34]
CGAP database http://cgap.nci.nih.gov	Database of cancer-related microarray datasets.	[34]
Oncomine http://www.oncomine.org	Database of cancer-related microarray datasets.	[34]
UniHI http://www.unihi.org	Database of human molecular interaction networks.	[34]
Pathguide http://www.pathguide.org	Database of 702 biological pathway related resources and molecular interactions.	[34]
UniProt http://www.uniprot.org	Encompassing protein information center.	[34]
InterPro http://www.ebi.ac.uk/interpro	Database of protein domain information.	[34]

Table 2. Web-tools and software utilized in target discovery.

Web-Tools/Software Used for Target Discovery	Specific Information	Ref.
GoPubMed http://www.gopubmed.org	PubMed search engine utilized as a text-mining tool.	[34]
Textpresso http://www.textpresso.org	Full-text engine used in text mining, classification, and search.	[34]
BioRAT http://bioinfadmin.cs.ucl.ac.uk/biorat/docs/index	Full-text search engine used for text mining.	[34]
ABNER http://pages.cs.wisc.edu/~bsettles/abner	Molecular biology text analyzer and entity tagger tool.	[34]
PPICurator https://ppicurator.hupo.org.cn	Tool used for mining comprehensive protein-protein interaction.	[34]
GeneWays http://geneways.genomeleft.columbia.edu	Biological pathway extracting tool.	[34]

Table 3. Databases used for lead discovery, optimization, and synthesis.

Database	Specific Information	Ref.
ADReCS http://bioinf.xmu.edu.cn/ADReCS	Database of toxicology information with 137,619 Drug-ADR pairs.	[35]
ChEMBL https://www.ebi.ac.uk/chembl	Database of drug-like small molecules with predicated bioactive properties.	[35]
ChemSpider http://www.chemspider.com	Encompassing database of over 64 million chemical structures.	[35]
DrugCentral http://drugcentral.org	Database containing relevant drug information of activity, chemical identity, mode of action, etc.	[35]

3. Random Forest (RF)

RF is a widely used algorithm explicitly designed for large datasets with multiple features, as it simplifies by removing outliers, as well as classify and designate datasets based on relative features classified for the particular algorithm. It is commonly trained for large inputs and variables and accessibility based on data collection from multiple databases. It is beneficial in different aspects, such as attributing missing data, working with outliers, and estimating characteristics for classification [25]. The underlying mathematical process of RF consists of several uncorrelated decision trees as an ensemble; each tree is responsible for determining one prediction. The one that constitutes the most votes is considered the best fit (Figure 2a) [36]. Although false positives may happen in any statistical analysis, RF, along with SVM and NB, has been suggested to make the least amount of errors compared to other algorithms. With multiple decision trees, individual errors are minimized due to their assemblies of several predictions rather than solely focusing on one prediction.

In drug discovery, RFs are mainly used either for feature selections, classifiers, or regression. Cano et al. utilized RF methods to improve affinity prediction between ligand and the protein by virtual screening through selecting molecular descriptors, based on a training data set for enzymes, such as ligands of kinases and nuclear hormone receptors. Some of the essential factors accompanying RF in drug discovery are: It expedites the training process, uses fewer parameters, imputes missing data, and incorporates nonparametric data [37]. Rahman et al. utilized multivariate RF by including information relating to genomic sequencing, which helped sustain error and achieve drug responses based on genomic characterizations. Multivariate RFs specialize in limiting error by calculating several error estimates techniques within the system. The computational framework inputs the data that incorporates genetic and epigenetic characterization combinations, allowing the framework to predict the mean and confidence interval of the drug responses. An important quality essential for analyzing any drug to be processed in clinical trials [38]. Rahman et al. endeavored to combine

the modeling framework with functional RF for improving the prediction based on the response profile. They tried to combat the difficulties observed in individuals related to finding appropriate compounds depending on individual tumors. RF was incorporated for the generation of the regression tree node and leaf nodes. It acquired the data points of dose-responses. The leaf nodes in the algorithms are responsible for making predictions about the dose-response profile, simultaneously storing the functional data. The model recorded data is comprised of the genome sequences and their characteristics [39]. RF algorithms have also been implemented as a method of classification and regression in a quantitative structure-activity relationship (QSAR) modeling used in lead discovery processes [40,41].

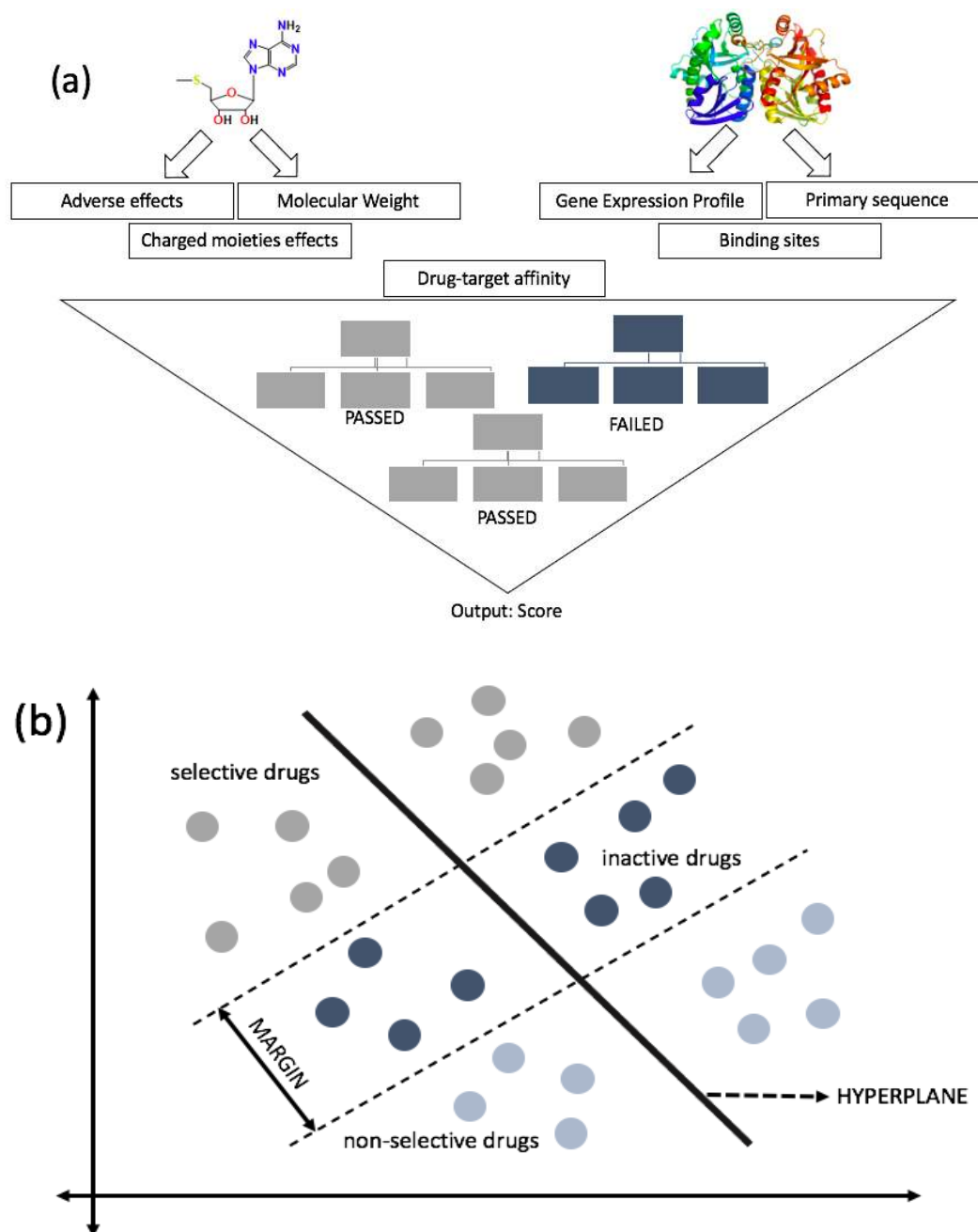


Figure 2. Schematic view of drug development using random forest (RF) (a) and support vector machine (SVM) (b). (a) RF reaches the final decision of drugs by combining the results of randomly-created decision trees (three trees are shown for simplicity). There are multiple features that the computational

queries look for in both target and drug. When there is a compatibility match, it proceeds to the next step to match additional features. A series of datasets is inputted into the query, and each tree is responsible for computing a prediction. The prediction picked by most trees is used for the next step. The system of using many decision trees is intended to minimize errors mathematically. (b) SVM utilizes similarities between the classes, called support vectors, to distinguish between the classes based on the trained features. It formulates hyperplanes that separate two classes (can be multiclass, if needed). SVM incorporates multiple training sets depending on the classifiers and formulates compounds' status (active or inactive). During the process, compounds are separated into three sections: Non-selective compounds (active), selective compounds (active), and in the margin are inactive compounds. Although non-selective compounds are active, they are not selective towards the protein of interest. In contrast, selective compounds are active and selective towards the protein of interest.

4. Naive Bayesian (NB)

NB algorithms are a subset of supervised learning methods that have become an essential tool used in predictive modeling classification. Standard NB algorithms work to classify features of datasets, and depending on the input characteristics, factor correlation, and dimensionality of the data, it can be one of the most efficient techniques for the task [42–44]. The effectiveness of NB alongside decision tree algorithms for the use of text mining has not been determined. These techniques enhance the accuracy of retrieved data sets, which generally originate in large, muddled sources [45,46]. Classification of biomedical data is crucial in the drug discovery process, especially in the target discovery subset. NB algorithms have shown great promise as classification tools for biomedical data, often filled with non-related information and data, known as noise [47]. NB techniques could also serve important roles in predicting ligand-target interactions, which could be a massive step forward in lead discovery [48]. Recently, researchers have been able to incorporate NB techniques into diverse applications within the drug discovery process. In a study, Pang et al. used NB models and additional techniques as classifiers for active and inactive compounds, with possible activity as antagonists for estrogen receptors in breast cancer [49]. The researchers utilized the ability of NB algorithms to process vast quantities of information while having a unique tolerance to random noise. The technique, in combination with other tools such as extended-connectivity fingerprint-6, was able to collect excellent outputs. In a recent study, Wei et al. utilized a combinational technique of NB and support vector machine algorithms to predict possible compounds that could be active against targets of human immunodeficiency virus type-1 and the hepatitis C virus generated from multiple QSAR models [50]. Their model utilized NB as a classifier technique paired alongside two different descriptor systems, one also being extended-connectivity fingerprint-6. The utilization of NB, combined with other systems and techniques, has shown to be useful in incorporating drug discovery processes.

5. Support Vector Machine (SVM)

SVMs are supervised machine learning algorithms used in drug discovery to separate classes of compounds based on the feature selector by deriving a hyperplane. It utilizes the similarities between classes to formulate infinite numbers of the hyperplane. For linear data, it trains by separating classes consisting of compounds based on selected features and projects them into chemical feature space. An optimal hyperplane attained by maximizing margin between classes in N-dimensional space (N is the number of features); it is denoted by a hyperplane, which is used to classify data points by setting decision boundaries [51]. SVM is crucial to drug discovery because of its capability of distinguishing between active and inactive compounds, ranking compounds from each database (shown in Figure 2b), or training regression model. Regression models are vital in determining the relationship between the drug and ligand, as it employs a query for datasets to predict [52–55]. When several active compounds are screened against a single protein of interest, SVM can be attributed in various scenarios. SVM classification has a subset binary class prediction that could differentiate between active from inactive molecules.

For drug discovery, it could rank compounds from different databases based on the probability of being active for any computational screening. SVM can be extrapolated in different ways to attain results, with a main focus to distinguish between active and inactive compounds. The process could be manipulated by training the algorithm using various descriptors for feature selectors such as 2D fingerprints, and target protein. A class label is formulated, negative or positive, depending on the direction where the compound is placed from the hyperplane, thereby ranking compounds from the most selective to least selective [55,56]. However, for non-linear data, kernel functions are utilized to optimize the results. Kernel functions plot the data in a higher-dimensional space, where the separation between classes is feasible.

For drug-target interaction, it is specifically designed for integrating ligands and proteins of interest information as an essential component for SVM modeling [51]. Wang et al. investigated drug-target interactions and integrated information obtained from published research of various source to enhance the prediction. They used kernel function to incorporate information on drug pharmacological and therapeutic effects, drug chemical structures, and protein genomic information to characterize the drug-target interactions. Generally, results from the different sources were all promising, and kernel function for the prediction of pharmacological and therapeutic effects displayed the most potential [57]. SVM are also frequently used in predicting drugs that could have multiple bioactivities. For example, Kawaii et al. used SVM classifiers to construct a query where drugs were set against hundreds of targets to establish different biological pathways targeting their bioactivities [58]. In another study, a similar process was used to determine the bioactivities for antihypertensive drugs. The information about the drug activity was obtained from the Market Driven Demand Response database, and a multi-label SVM was employed to produce the query that shows the bioanalysis of drugs [59,60]. Drugs were discovered to be dual inhibitors against both angiotensin-converting enzyme I and neutral endopeptidases.

6. Limitations

ML algorithms have been an essential component of drug discovery. These methods increase efficiency and explore thousands of combinations that would have been impossible without this technology. As stated earlier, algorithms are trained with inputted data, but there are a few constraints with this technique. Although ML has been around for quite some time now, the biological pathways/targets being discovered are still novel. Information for the particular protein of interest might be limited, resulting in not much-extrapolated data. Free Energy Perturbation method is a platform where biological information regarding the protein is generated based on computational screening [61]. Data gathered from this method is utilized for training algorithms; however, not all the information is collected from a wet lab, rather computer-generated prediction is utilized. The accuracy of the training data might be lower than anticipated. Even though algorithms discussed in this review have a higher threshold for minimizing errors, there are still some categorical errors from training sets [61].

A more concise way to understand this is by the statistical angle. With algorithms prediction, there is always a concern with overfitting or underfitting. Overfitting is when the model consists of lower quality information/technique but generates higher quality performance. It occurs when the model picks up unusual features during the training, resulting in a negative impact on the model [17]. In contrast, underfitting models fail to recognize the data sets' underlying trend and generalize the new data inputted [62]. Both underfitting and overfitting result in inaccurate results. There are several ways to tackle overfitting and underfitting, such as increasing the sample size and cross-validation. Cross validation is an often-used technique used to estimate the accuracy of the ML algorithms' models, by using independent data sets to infer the models.

Another concern raised by chem-informaticians is ample chemical space constructed through algorithms [52,63]. The chemical area is a relative set of descriptors, consisting of thousands of compounds within a frame with boundaries generated by ML algorithms [64,65]. The challenge with

chemical space is the clustering of compounds with high density, which often leads to avoidance of compounds with some essential properties and compounds. Studies regarding these issues are discussed later, models to augment chemical space coverage to highlight the molecules with properties different from others [19,66].

7. Deep Learning (DL) Methods

DL algorithms are considered one of the cutting-edge areas of development and study in almost all scientific and technological fields. The renaissance of artificial NNs into workable algorithms from their former theorized and predicted applications, first developed in the 1950s, is an essential pillar of DL and the continued success brought by AI-based integration of standard techniques. DL algorithms give computational models the ability to learn a representation of multidimensional data through abstraction [67]. DL has allowed for resolving many challenges faced by standard ML algorithms, including image recognition and speech recognition. In the drug discovery process, DL techniques have become exemplar methods of drug activity prediction, target discovery, and lead molecule discovery [68–70]. The basis of DL is often implicated in NN systems, where they are used to create systems that have the capability to complete complex data recognition, interpretation, and generation. The main subsets of artificial NNs used in current drug discovery are deep neural networks (DNNs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs).

The utilization of specific NNs from the variations that exist in the subset is dependent on multiple factors. DNNs, a specific type of feed forward neural networks, function with singular path data flow from the input layer through the hidden layer(s) reaching an output layer (Figure 3a). The outputs generated are typically identified using trained supervised learning algorithms. DL algorithms function through neural networks which can incorporate other ML techniques for training. Through supervised and reinforcement learning guided methods, a DNN can be trained to complete complex tasks. A generative DNN can create novel chemical compounds from existing libraries and training sets (Figure 3a); while, a predictive DNN can predict the chemical attributes of the novel compounds [71,72]. QSAR models are currently being used to find the correlation between these compounds' chemical structure and activity. QSAR analysis is one of the most advanced forms of DL-based AI in current drug discovery and development. It has allowed researchers to take 2D chemical structures and determine physicochemical descriptors related to the molecule's activity. 3D-QSAR has allowed further inquiry of geometric structure impacting ligand-target interactions [33,73,74]. QSAR has been actively used in the pharmaceutical industry to predict on/off-target activities of developed lead compounds on specific targets. These algorithmic approaches to discovery and development are not, by all means, full proof or thoroughly capable.

There are always some error sources and imprecision over the multiplicity of studies conducted using these AI algorithms. It has been found that NNs face a few deficiencies in comparison to other ML algorithms in their applications of QSAR studies. The first being the presence of excess descriptors that cause redundancy in NN and eventual clogging of outputs. This redundancy can significantly drop the efficiency of the NN, while also creating non-ideal outputs. Unknown descriptors also pose an issue because they may also affect the output generated. These issues have been alleviated using more specific feature selection algorithms to get a smaller number of higher quality descriptors; however, it will continue to be a problem faced by NN-based QSAR. The second issue with these NN-based assays is implementing ideal network parameters without overfitting [74]. Remedies to this issue have been proposed and implemented, but it persists to be a recurring issue without the necessary adjustments [75].

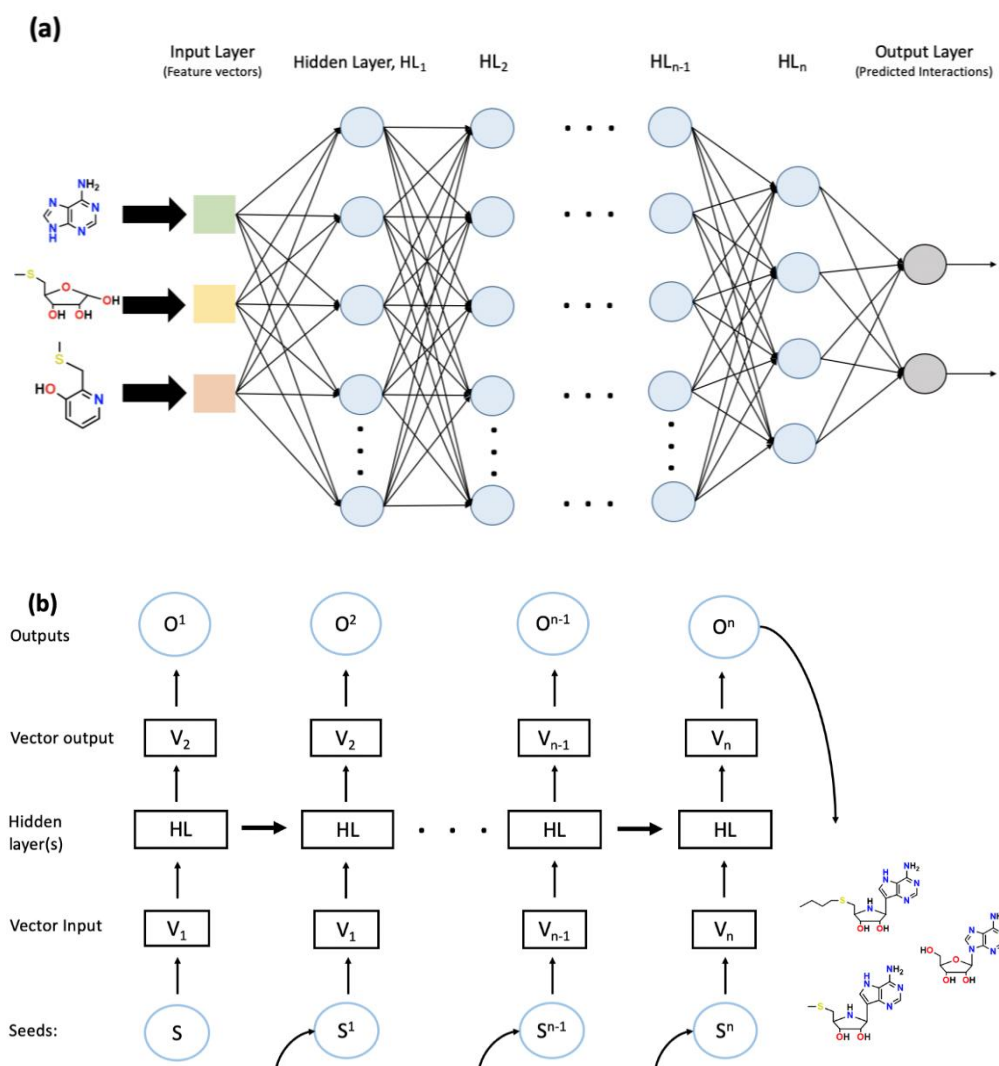


Figure 3. The general scheme of deep neural network (DNN) (a) and recurrent neural network (RNN) (b). (a) DNN consists of an input layer followed by several hidden layers and an output layer. In this case, the input layer utilizes feature vectors generated by a convolutional network. The progression of the NN follows a single path through hidden layer 1 (HL_1) to HL_n , indicating the feedforward nature of the NN. The generated outputs are often processed using supervised learning techniques for the identification and collection of sensible interactions. (b) RNN begins with a seed, S , which is inputted into the system. Through the use of algorithmic processing, the seed is turned into a reference vector, V_1 , which is used by the HL to generate a vector output, V_2 . V_2 is subsequently optimized through input training sets and creates the output, O . The generation of these outputs eventually leads to the creation of a gatherable data set. In the meantime, the HLs feed forward to provide information from previous steps. One example is chemical structure generation using SMILE string characters as seeds; hence the desired gathered outputs would be a string of SMILE characters that would be the desired structure. The dataset created in the figure is gathered and analyzed into the resultant molecules.

Once the initial work of target discovery is complete and better understanding is developed for target-molecule interaction, chemical synthesis and characterization become a priority in the pipeline. An important note in this process is using descriptive simplified molecular-input line-entry system (SMILES) nomenclature in much of the algorithms regarding de novo drug design and discovery. RNNs, which are a type of NN that utilize a system of self-learning through generational processing of the inputs and developing hidden layers. The subset RNN-type long short-term memory have become a reliable, standardized method for generating novel chemical structures. RNNs are unique

in their ability to use neurons connected in the same hidden layer to form a functioning cycle of processing inputs and outputs compared to DNNs and feedforward neural networks (Figure 3b), which have no connections within the same layer and only push outputs. These generative RNNs have shown promising results in the generation of sensible, structurally correct, and feasible, novel SMILE structures that were not included in the original SMILE training sets [76–79]. A recent study by Segler et al. used generative RNN models to develop possible molecular structures that could have activity against *Staphylococcus aureus* (*S. aureus*) and *Plasmodium falciparum* (*P. falciparum*). Their models were given small sets of molecular structures that had known activity against these target organisms; from these inputs, the model generated 14% of the 6051 potential molecule candidates for *S. aureus* that has been developed by medicinal chemists. The model also generated 28% of the existing compounds developed for *P. falciparum* [80]. Traditionally, the generation and implementation of chemical synthesis routes have been the sole responsibility of chemists. However, this role is evolving to include more and more computational based synthesis, also known as computer-aided synthesis planning (CASP), with the emergence of AI [81–83]. The Monte Carlo tree search (MCTS) based through NN techniques have been used in current studies to generate CASP workflows. The MCTS technique is ideal for this purpose because of the simulation's ability to perform random continuous step searches without branching until optimal conditions and solutions are met [82,83]. In a groundbreaking study conducted by Segler and Waller [84], an MCTS method using three NNs alongside 12.4 million transformation rules, retrieved through AI-based data mining, from all the available chemical synthesis literature at the time to generate a sensible workflow for CASP. The first NN, an expansion node, retrospectively searches for new transformations to create the molecule; it also predicts the feasibility of applying the transformation from the 12.4 million transformation rules. This allows the expansion node to select the best, as in most feasible and high yielding, transformations from the literature. The second NN, a rollout node, filters the inputs to include only the most frequently reported transformation rules to enable the best possibilities of successful transformations. The update node then incorporates the new pathway into the search tree. This algorithm was able to solve 80% of retrosynthesis problems in just 5 s, and >90% of problems in 60 s [82–84]. Various studies have been conducted to optimize AI-based chemical synthesis and reaction routes [85–87]. Through the further implementation of AI-based chemical synthesis and characterization, it will be possible to move drug discovery further from the bench to in silico and increase the time and cost-efficiency of discovery and development.

CNNs are a subset of DNNs that take inputs, assign weights to specific parts of the input, then build the ability to differentiate the data. While traditional DNNs are limited in their ability to function correctly on higher-dimensional datasets, CNNs serve as a gleaming solution to tackling this issue with their ability to preserve input dimensionality. The training required for a CNN model is significantly less than DNNs, and RNNs would need to function with reasonable accuracy and efficacy. These advantages have allowed it to become a prominent learning algorithm for image recognition, surpassing other standard ML algorithms. In the process of drug discovery, CNNs have become efficient tools used in target discovery, lead discovery and characterization, in silico target-lead interaction screening, and protein-ligand scoring [68,88–90]. Combinations of these DL techniques, such as CNNs, have also been very successful in identifying gene mutations and disease targets [91,92]. The incorporation of CNNs into drug development is not merely limited to target discovery; it has also been widely used in later-stage development. One such use of CNNs in this manner to assist in the generation of motility models of cancer cells responding to treatment [93]. In a recent study, Feng, Zhang, and Shi demonstrated the use of deep learning based drug-drug interaction (DDI) predictors [94], with the aim to address a wet lab issue during the drug discovery, which is often costly and time consuming. The researchers developed a new method utilizing graph convolutional networks and DNN models. In their design, the graph convolutional network served as a structure feature extractor from drugs found in DDI, learning low-dimensional representations (vectors) of the features from the DDI networks. The information is then taken to the DDN model which served as the actual predictor; the ability of the model to take the feature vectors and link them with corresponding feature

vectors of possible drug combinations allowed it to produce the interaction prediction. Encouragingly, the predictions using their method outclassed popularly used state-of-the-art-methods [94].

8. Examples of Drug Discovery (Paper Summaries and Relevance to Topic)

ML is already being used to develop novel molecules that could be used as future antibiotic candidates. In a recent, groundbreaking study conducted by Stokes et al., the researchers demonstrated the utility and capability of ML techniques in the drug discovery process [95]. They specifically capitalized on the use of DNNs to create novel molecules with broad-spectrum antibacterial activity. These discovered candidates were also identified to be structurally distinct from any known antibiotics. The researchers utilized a training set of 2335 molecules for a DNN model to predict the growth inhibition of *Escherichia coli*, followed by the running of the model on greater than 107 million molecules from several chemical libraries. This gave the researchers the ability to identify potential lead compound candidates that may have similar bioactivity. Through scoring generated by the model, the researchers were able to identify a list of sensible candidates that meet a predetermined score threshold and various other eliminative criteria. The researchers' efforts proved fruitful, and they were able to identify a c-Jun N-terminal kinase inhibitor, halicin, that is distinct from known antibiotics. This antibacterial candidate was also discovered to be a potent growth inhibitor of *Escherichia coli*, and had shown efficacy against *Clostridioides difficile* and *Acinetobacter baumannii* infections in murine models [95]. In a study conducted by Fields et al., ML algorithms, including NNs-based techniques and SVM models, were used to discover novel antimicrobial peptides, also known as bacteriocins, from bacteria could ultimately be used as compelling antibiotic candidates [96]. Discoveries such as that of the bacteriocins are the outcomes of the machine-learning algorithm's ability to build and function as complex processing systems. In the study, a positive and negative training set of 346 bacteriocins was used to train the algorithm. These input bacteriocins were represented as complicated vector sums. The machine-learning algorithm then took the inputs and generated new vector structure outputs that preserved the original inputs' key features. These outputs were translated into 676 bacteriocins that were not identical to the input bacteriocins. From the output bacteriocins, 28,895 peptides were generated using a sliding window algorithm; these peptides spanned 20-mers and were placed through biophysical parameters. Fields et al. then selected 16 of the highest affinity peptides from the biophysical filtration for in vitro testing. Their finding indicated that the peptides had significant antimicrobial activity against *Escherichia coli* and *Pseudomonas aeruginosa* [96].

The utility of ML-based mining has proved to be extremely advantageous with the advent of high throughput data generation and collection. These algorithms have been extensively used alongside the vast data generated utilizing high-throughput sequencing to enhance the target discovery process [15,97]. The innovation of algorithm-assisted data collection and manipulation has already been implemented in emerging research; recently, it has been used to find novel molecular therapeutic targets for aggressive melanoma. Researchers were able to use unsupervised learning techniques through GeneCluster to identify groups of cell lines, one was a primary melanoma group, and the other was an aggressive melanoma group. Through further analysis using supervised learning techniques, the researchers were able to identify invasion-specific genes related to aggressive melanoma [98]. One of the many challenges with cancer treatments is detecting response profiles designed primarily for individual patients. Sakellaropoulos et al. built a network-based framework. They trained a database containing 1001 cancer cell lines, from the Genomics of Drug Sensitivity in Cancer, using DNN to predict drug responses based on gene expressions. The results were evaluated in several clinical cohorts. DNNs are observed to outperform several others in silico screening due to their capability to embrace biological interactions and create models that can capture the biological complexities and accurately predict clinical response with the help of cancer cell baselines. Their model incorporated RF and elastic net (Enet) algorithms to evaluate the DNN model's results. This framework was only tested on five patients; thus, not much coverage was obtained through this model; therefore, they expanded their study to a more massive sample size. They utilized response data for two drugs: Cisplatin and

paclitaxel, and analyzed it with gene expression profiles and patients' responses to those two drugs gathered from different clinical trials. The study was done on a small scale, implementing DL network training sets and ML algorithms, with a limited amount of knowledge. It is believed that ML could essentially be a powerful tool to assist within the medicinal field, as more data and information are retrieved on patient response profiles [99].

The diseases discussed have been around for a long time, but the emergent need for a treatment for Coronavirus disease 2019 (COVID-19) has stirred up the research world. The pandemic outbreak has caused detrimental effects around the world, but the COVID-19 virus (SARS-CoV-2) is a novel strain of the same species of virus causing the 2003 Severe acute respiratory syndrome (SARS-CoV-1); thus, several studies are incorporating earlier information into supervised ML to quickly find a remedy for this virus [100]. Researchers worldwide are exhausting all available resources, and ML has helped narrow down the drug candidates and minimize clinical trial failure. Kowalewski and Ray developed ML models to help identify effective drugs against 65 human proteins (target) studied to interact with SARS-CoV-2 proteins. As the virus is known to target the respiratory tract, including nasal epithelial cells and upper airway and lungs, they deduce it from inhaling therapeutics to directly target the damaged cells. They assembled 14 million chemicals from ZINC databases and utilized ML models to predict vapor pressure and mammalian toxicity to rank the chemicals and find drugs that share the same chemical space. Their main goal was to establish a short term and long-term pipeline for future purposes. They utilized SVM and RF to create models that could predict drugs and their efficacy. Although most of the researchers focus on a single protein responsible for replication and host entry, it might only allow short term repair. In the long term, Kowalewski and Ray proposed to look into multiple drugs that could potentially target various proteins with diverse biological pathways [101].

9. Conclusions

ML-based techniques seek to revitalize the development of drugs. These methods are based on separate applications in target discovery, lead compound discovery, synthesis, protein-ligand interactions, etc. ML applications are paving the way for algorithm-enhanced data query, analysis, and generation. One such example is ML incorporated into target discovery, based heavily on the refinement and search of existing omics and medical data. Through AI integration using ML techniques, viable targets can be found using data clustering, regression, and classification from vast omics databases and sources. Lead compound discovery, e.g., using QSAR, is currently frequently used to develop sensible molecular candidates based on training inputs. Lead compound synthesis has also been expedited with NN-based retrosynthesis algorithms alongside best-chance trees with the input of vast amounts of accumulated data and rules to develop algorithms that can generate synthesis pathways with greater than 90% accuracy in 60 s. Applications of ML in the processes of drug development have been used for some time now. These applications have proven to be steps above previous methods; the development of ML and DL techniques are not all brand new. They have been carefully crafted and developed through decades of research. This curation of function and utility to ML algorithms and techniques has allowed for the continued success and development in drug discovery. Owing to more precise algorithms, more powerful supercomputers, and substantial private and public investment into the field, these applications are becoming more intelligent, cost-effective, and time-efficient while boosting efficacy.

Funding: This work is supported in part by the National Science Foundation under Award No. OIA-1946391 and by Arkansas Division of Higher Education under 2019–2020 SURF; DWU is funded in part by the Arkansas Research Alliance.

Conflicts of Interest: The authors declare no conflict of interest; the funders had no role in the study design, nor in collection, analysis, or interpretation of the data. The funders had no role in the writing of the manuscript or in the decision to publish the results.

References

1. Fayyad, J.; Jaradat, M.A.; Gruyer, D.; Najjaran, H. Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review. *Sensors* **2020**, *20*, 4220. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Deng, L.; Li, X. Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1060–1089. [\[CrossRef\]](#)
3. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep Audio-visual Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *1*. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Joachims, T.; Radlinski, F. Search Engines that Learn from Implicit Feedback. *Computer* **2007**, *40*, 34–40. [\[CrossRef\]](#)
5. Morgan, S.; Grootendorst, P.; Lexchin, J.; Cunningham, C.; Greyson, D. The cost of drug development: A systematic review. *Health Policy* **2011**, *100*, 4–17. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Ng, H.W.; Zhang, W.; Shu, M.; Luo, H.; Ge, W.; Perkins, R.; Tong, W.; Hong, H. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. *BMC Bioinform.* **2014**, *15*, S4. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Ng, H.W.; Shu, M.; Luo, H.; Ye, H.; Ge, W.; Perkins, R.; Tong, W.; Hong, H. Estrogenic activity data extraction and in silico prediction show the endocrine disruption potential of bisphenol A replacement compounds. *Chem. Res. Toxicol.* **2015**, *28*, 1784–1795. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Hong, H.; Neamati, N.; Winslow, H.E.; Christensen, J.L.; Orr, A.; Pommier, Y.; Milne, G.W.A. Identification of HIV-1 integrase inhibitors based on a four-point pharmacophore. *Antivir. Chem. Chemother.* **1998**, *9*, 461–472. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Hong, H.; Tong, W.; Xie, Q.; Fang, H.; Perkins, R. An in silico ensemble method for lead discovery: Decision forest. *SAR QSAR Environ. Res.* **2005**, *16*, 339–347. [\[CrossRef\]](#)
10. Hong, H.; Fang, H.; Xie, Q.; Perkins, R.; Sheehan, D.M.; Tong, W. Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. *SAR QSAR Environ. Res.* **2003**, *14*, 373–388. [\[CrossRef\]](#)
11. Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Talevi, A.; Morales, J.F.; Hather, G.; Podichetty, J.T.; Kim, S.; Bloomingdale, P.C.; Kim, S.; Burton, J.; Brown, J.D.; Winterstein, A.G.; et al. Machine Learning in Drug Discovery and Development Part 1: A Primer. *CPT Pharmacomet. Syst. Pharmacol.* **2020**, *9*, 129–142. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Gertrudes, J.C.; Maltarollo, V.G.; Silva, R.A.; Oliveira, P.R.; Honório, K.M.; da Silva, A.B. Machine learning techniques and drug design. *Curr. Med. Chem.* **2012**, *19*, 4289–4297. [\[CrossRef\]](#)
14. Agarwal, S.; Dugar, D.; Sengupta, S. Ranking chemical structures for drug discovery: A new machine learning approach. *J. Chem. Inf. Model.* **2010**, *50*, 716–731. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Rodrigues, T.; Bernardes, G.J.L. Machine learning for target discovery in drug development. *Curr. Opin. Chem. Biol.* **2020**, *56*, 16–22. [\[CrossRef\]](#)
16. Gao, D.; Chen, Q.; Zeng, Y.; Jiang, M.; Zhang, Y. Application of Machine Learning on Drug Target Discovery. *Curr. Drug Metab.* **2020**. [\[CrossRef\]](#)
17. 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 Discov.* **2019**, *18*, 463–477. [\[CrossRef\]](#)
18. Zoffmann, S.; Vercruysse, M.; Benmansour, F.; Maunz, A.; Wolf, L.; Marti, R.B.; Heckel, T.; Ding, H.; Truong, H.H.; Prummer, M.; et al. Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **2019**, *9*, 5013. [\[CrossRef\]](#)
19. Ekins, S.; Puhl, A.C.; Zorn, K.M.; Lane, T.R.; Russo, D.P.; Klein, J.J.; Hickey, A.J.; Clark, A.M. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **2019**, *18*, 435–441. [\[CrossRef\]](#)
20. Khamis, M.A.; Gomaa, W.; Ahmed, W.F. Machine learning in computational docking. *Artif. Intell. Med.* **2015**, *63*, 135–152. [\[CrossRef\]](#)
21. Leelananda, S.P.; Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **2016**, *12*, 2694–2718. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Maia, E.H.B.; Assis, L.C.; de Oliveira, T.A.; da Silva, A.M.; Taranto, A.G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **2020**, *8*, 343. [\[CrossRef\]](#) [\[PubMed\]](#)

23. Talambedu, U.; Shanmugarajan, D.; Goyal, A.K.; Kumar, C.S.; Middha, S.K. Recent Updates on Computer-aided Drug Discovery: Time for a Paradigm Shift. *Curr. Top. Med. Chem.* **2017**, *17*, 3296–3307. [\[CrossRef\]](#)
24. Réda, C.; Kaufmann, E.; Delahaye-Duriez, A. Machine learning applications in drug development. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 241–252. [\[CrossRef\]](#)
25. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
26. Webb, G.I. Naïve Bayes. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2010. [\[CrossRef\]](#)
27. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
28. Rifaioglu, A.S.; Atas, H.; Martin, M.J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. *Brief Bioinform.* **2019**, *20*, 1878–1912. [\[CrossRef\]](#)
29. Dugger, S.A.; Platt, A.; Goldstein, D.B. Drug development in the era of precision medicine. *Nat. Rev. Drug Discov.* **2018**, *17*, 183–196. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafner, D.A.; McKinney, E.F. From Big Data to Precision Medicine. *Front. Med.* **2019**, *6*, 34. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Liu, B.; He, H.; Luo, H.; Zhang, T.; Jiang, J. Artificial intelligence and big data facilitated targeted drug discovery. *Stroke Vasc. Neurol.* **2019**, *4*, 206. [\[CrossRef\]](#)
32. Cirillo, D.; Valencia, A. Big data analytics for personalized medicine. *Curr. Opin. Biotechnol.* **2019**, *58*, 161–167. [\[CrossRef\]](#)
33. Chen, R.; Liu, X.; Jin, S.; Lin, J.; Liu, J. Machine Learning for Drug-Target Interaction Prediction. *Molecules* **2018**, *23*, 2208. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Yang, Y.; Adelstein, S.J.; Kassis, A.I. Target discovery from data mining approaches. *Drug Discov. Today* **2009**, *14*, 147–154. [\[CrossRef\]](#)
35. Yella, J.K.; Yaddanapudi, S.; Wang, Y.; Jegga, A.G. Changing Trends in Computational Drug Repositioning. *Pharmaceuticals* **2018**, *11*, 57. [\[CrossRef\]](#)
36. Sarica, A.; Cerasa, A.; Quattrone, A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer’s Disease: A Systematic Review. *Front. Aging Neurosci.* **2017**, *9*, 329. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Cano, G.; Garcia-Rodriguez, J.; Garcia-Garcia, A.; Perez-Sanchez, H.; Benediktsson, J.; Thapa, A.; Barr, A. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Syst. Appl.* **2017**, *72*, 151–159. [\[CrossRef\]](#)
38. Rahman, R.; Otridge, J.; Pal, R. IntegratedMRF: Random forest-based framework for integrating prediction from different data types. *Bioinformatics* **2017**, *33*, 1407–1410. [\[CrossRef\]](#)
39. Rahman, R.; Dhruva, S.R.; Ghosh, S.; Pal, R. Functional random forest with applications in dose-response predictions. *Sci. Rep.* **2019**, *9*, 1628. [\[CrossRef\]](#)
40. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [\[CrossRef\]](#)
41. Lee, K.; Lee, M.; Kim, D. Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server. *BMC Bioinform.* **2017**, *18*, 567. [\[CrossRef\]](#)
42. Bielza, C.; Larrañaga, P. Discrete Bayesian Network Classifiers: A Survey. *ACM Comput. Surv.* **2014**, *47*, 43. [\[CrossRef\]](#)
43. Gilboa, E.; Saatçi, Y.; Cunningham, J.P. Scaling Multidimensional Inference for Structured Gaussian Processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Sun, H. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48*, 4031–4039. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Ratanamahatana, C.A.; Gunopulos, D. Feature selection for the naive bayesian classifier using decision trees. *Appl. Artif. Intell.* **2010**. [\[CrossRef\]](#)
46. Kim, S.-B.; Han, K.-S.; Rim, H.-C.; Myaeng, S.-H. Some Effective Techniques for Naive Bayes Text Classification. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1457–1466. [\[CrossRef\]](#)
47. Anagaw, A.; Chang, Y.-L. A new complement naïve Bayesian approach for biomedical data classification. *J. Ambient Intell. Hum. Comput.* **2019**, *10*, 3889–3897. [\[CrossRef\]](#)

48. Nigsch, F.; Bender, A.; Jenkins, J.L.; Mitchell, J.B.O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313–2325. [\[CrossRef\]](#)
49. Pang, X.; Fu, W.; Wang, J.; Kang, D.; Xu, L.; Zhao, Y.; Liu, A.L.; Du, G.H. Identification of Estrogen Receptor α Antagonists from Natural Products via In Vitro and In Silico Approaches. *Oxid. Med. Cell. Longev.* **2018**, *2018*, 6040149. [\[CrossRef\]](#)
50. Wei, Y.; Li, W.; Du, T.; Hong, Z.; Lin, J. Targeting HIV/HCV Coinfection Using a Machine Learning-Based Multiple Quantitative Structure-Activity Relationships (Multiple QSAR) Method. *Int. J. Mol. Sci.* **2019**, *20*, 3572. [\[CrossRef\]](#)
51. Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. *Expert Opin. Drug Discov.* **2014**, *9*, 93–104. [\[CrossRef\]](#)
52. Maltarollo, V.G.; Kronenberger, T.; Espinoza, G.Z.; Oliveira, P.R.; Honorio, K.M. Advances with support vector machines for novel drug discovery. *Expert Opin. Drug Discov.* **2019**, *14*, 23–33. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Lima, A.N.; Philot, E.A.; Trossini, G.H.; Scott, L.P.; Maltarollo, V.G.; Honorio, K.M. Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* **2016**, *11*, 225–239. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Korkmaz, S.; Zararsiz, G.; Goksuluk, D. Drug/nondrug classification using Support Vector Machines with various feature selection strategies. *Comput. Methods Programs Biomed.* **2014**, *117*, 51–60. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Wassermann, A.M.; Geppert, H.; Bajorath, J. Application of support vector machine-based ranking strategies to search for target-selective compounds. *Methods Mol. Biol.* **2011**, *672*, 517–530. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Ostermann, C.; Zell, A. Large-scale learning of structure-activity relationships using a linear support vector machine and problem-specific metrics. *J. Chem. Inf. Model.* **2011**, *51*, 203–213. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Wang, Y.C.; Zhang, C.H.; Deng, N.Y.; Wang, Y. Kernel-based data fusion improves the drug-protein interaction prediction. *Comput. Biol. Chem.* **2011**, *35*, 353–362. [\[CrossRef\]](#)
58. Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive activity profiling of drugs by topological-fragment-spectra-based support vector machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152–1160. [\[CrossRef\]](#)
59. Kawai, K.; Takahashi, Y. Identification of the Dual Action Antihypertensive Drugs Using TFS-Based Support Vector Machines. *Chem. Bio Inform. J.* **2009**, *9*, 41–51. [\[CrossRef\]](#)
60. Rossi, G.P.; Dual, A.C.E. NEP inhibitors: A review of the pharmacological properties of MDL 100240. *Cardiovasc. Drug Rev.* **2003**, *21*, 51–66. [\[CrossRef\]](#)
61. Kaiser, T.M.; Burger, P.B. Error Tolerance of Machine Learning Algorithms across Contemporary Biological Targets. *Molecules* **2019**, *24*, 2115. [\[CrossRef\]](#)
62. Lever, J.; Krzywinski, M.; Altman, N. Model selection and overfitting. *Nat. Methods* **2016**, *13*, 703–704. [\[CrossRef\]](#)
63. Arús-Pous, J.; Awale, M.; Probst, D.; Reymond, J.L. Exploring Chemical Space with Machine Learning. *Chimia* **2019**, *73*, 1018–1023. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Von Lilienfeld, O.A.; Müller, K.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347–358. [\[CrossRef\]](#)
65. Dobson, C. Chemical space and biology. *Nature* **2004**, *432*, 824–828. [\[CrossRef\]](#) [\[PubMed\]](#)
66. Gromski, P.S.; Henson, A.B.; Granda, J.M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **2019**, *3*, 119–128. [\[CrossRef\]](#)
67. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
68. Dana, D.; Gadhiya, S.V.; St Surin, L.G.; Li, D.; Naaz, F.; Ali, Q.; Paka, L.; Yamin, M.A.; Narayan, M.; Goldberg, I.D. Deep Learning in Drug Discovery and Medicine; Scratching the Surface. *Molecules* **2018**, *23*, 2384. [\[CrossRef\]](#)
69. Korotcov, A.; Tkachenko, V.; Russo, D.P.; Ekins, S. Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, *14*, 4462–4475. [\[CrossRef\]](#)
70. Ekins, S. The Next Era: Deep Learning in Pharmaceutical Research. *Pharm. Res.* **2016**, *33*, 2594–2603. [\[CrossRef\]](#)
71. D'Souza, S.; Prema, K.V.; Balaji, S. Machine learning models for drug-target interactions: Current knowledge and future directions. *Drug Discov. Today* **2020**, *25*, 748–756. [\[CrossRef\]](#)

72. Baskin, I.I.; Winkler, D.; Tetko, I.V. A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* **2016**, *11*, 785–795. [[CrossRef](#)] [[PubMed](#)]
73. Verma, J.; Khedkar, V.M.; Coutinho, E.C. 3D-QSAR in drug design—A review. *Curr. Top. Med. Chem.* **2010**, *10*, 95–115. [[CrossRef](#)] [[PubMed](#)]
74. Ghasemi, F.; Mehridehnavi, A.; Pérez-Garrido, A.; Pérez-Sánchez, H. Neural network and deep-learning algorithms used in QSAR studies: Merits and drawbacks. *Drug Discov. Today* **2018**, *23*, 1784–1790. [[CrossRef](#)] [[PubMed](#)]
75. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
76. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [[CrossRef](#)] [[PubMed](#)]
77. Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885. [[CrossRef](#)] [[PubMed](#)]
78. Yuan, W.; Jiang, D.; Nambiar, D.K.; Liew, L.P.; Hay, M.P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.T.; Tibshirani, R.; et al. Chemical Space Mimicry for Drug Discovery. *J. Chem. Inf. Model.* **2017**, *57*, 875–882. [[CrossRef](#)]
79. Gupta, A.; Müller, A.T.; Huisman, B.J.H.; Fuchs, J.A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111. [[CrossRef](#)]
80. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131. [[CrossRef](#)]
81. De Almeida, A.F.; Moreira, R.; Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **2019**, *3*, 589–604. [[CrossRef](#)]
82. Chan, H.C.S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **2019**, *40*, 592–604. [[CrossRef](#)] [[PubMed](#)]
83. Segler, M.H.S.; Preuss, M.; Waller, M.P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610. [[CrossRef](#)] [[PubMed](#)]
84. Segler, M.H.S.; Waller, M.P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. Eur. J.* **2017**, *23*, 5966–5971. [[CrossRef](#)] [[PubMed](#)]
85. Zhou, Z.; Li, X.; Zare, R.N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337–1344. [[CrossRef](#)]
86. Coley, C.W.; Barzilay, R.; Jaakkola, T.S.; Green, W.H.; Jensen, K.F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443. [[CrossRef](#)]
87. Lee, A.A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J.L.; Butler, C.R. Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.* **2019**, *55*, 12152–12155. [[CrossRef](#)]
88. Reher, R.; Kim, H.W.; Zhang, C.; Mao, H.H.; Wang, M.; Nothias, L.F.; Caraballo-Rodriguez, A.M.; Glukhov, E.; Teke, B.; Leao, T.; et al. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J. Am. Chem. Soc.* **2020**, *142*, 4114–4120. [[CrossRef](#)]
89. Rath, P.C.; Ludlow, R.F.; Verdonk, M.L. Practical High-Quality Electrostatic Potential Surfaces for Drug Discovery Using a Graph-Convolutional Deep Neural Network. *J. Med. Chem.* **2020**, *63*, 8778–8790. [[CrossRef](#)]
90. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D.R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957. [[CrossRef](#)]
91. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [[CrossRef](#)]
92. Chang, P.; Grinband, J.; Weinberg, B.D.; Bardis, M.; Khy, M.; Cadena, G.; Su, M.Y.; Cha, S.; Filippi, C.G.; Bota, D.; et al. Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. *Am. J. Neuroradiol.* **2018**, *39*, 1201–1207. [[CrossRef](#)] [[PubMed](#)]
93. Mencattini, A.; di Giuseppe, D.; Comes, M.C.; Casti, P.; Corsi, F.; Bertani, F.R.; Ghibelli, L.; Businaro, L.; di Natale, C.; Parrini, M.C.; et al. Discovering the hidden messages within cell trajectories using a deep learning approach for in vitro evaluation of cancer drug treatments. *Sci. Rep.* **2020**, *10*, 7653. [[CrossRef](#)] [[PubMed](#)]

94. Feng, Y.-H.; Zhang, S.-W.; Shi, J.-Y. DPDDI: A deep predictor for drug-drug interactions. *BMC Bioinform.* **2020**, *21*, 419. [[CrossRef](#)]
95. Stokes, J.M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N.M.; MacNair, C.R.; French, S.; Carfrae, L.A.; Bloom-Ackerman, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702. [[CrossRef](#)] [[PubMed](#)]
96. Fields, F.R.; Freed, S.D.; Carothers, K.E.; Hamid, M.N.; Hammers, D.E.; Ross, J.N.; Kalwajtys, V.R.; Gonzalez, A.J.; Hildreth, A.D.; Friedberg, I.; et al. Novel antimicrobial peptide discovery using machine learning and biophysical selection of minimal bacteriocin domains. *Drug Dev. Res.* **2020**, *81*, 43–51. [[CrossRef](#)] [[PubMed](#)]
97. Reker, D.; Bernardes, G.J.L.; Rodrigues, T. Computational advances in combating colloidal aggregation in drug discovery. *Nat. Chem.* **2019**, *11*, 402–418. [[CrossRef](#)] [[PubMed](#)]
98. Ryu, B.; Kim, D.S.; DeLuca, A.M.; Alani, R.M. Comprehensive Expression Profiling of Tumor Cell Lines Identifies Molecular Signatures of Melanoma Progression. *PLoS ONE* **2007**, *2*, e594. [[CrossRef](#)]
99. Sakellaropoulos, T.; Vougas, K.; Narang, S.; Koinis, F.; Kotsinas, A.; Polyzos, A.; Moss, T.J.; Piha-Paul, S.; Zhou, H.; Kardala, E.; et al. A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Rep.* **2019**, *29*, 3367–3373.e4. [[CrossRef](#)]
100. Mohanty, S.; Rashid, M.H.A.; Mridul, M.; Mohanty, C.; Swayamsiddha, S. Application of Artificial Intelligence in COVID-19 drug repurposing. *Diabetes Metab. Syndr.* **2020**, *14*, 1027–1031. [[CrossRef](#)]
101. Kowalewski, J.; Ray, A. Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space. *Heliyon* **2020**, *6*, e04639. [[CrossRef](#)]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).