Received: 28 May 2021 | Revised: 13 July 2021 | Accepted: 16 July 2021
DOI: 10.1002/wcms.1568

ADVANCED REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship

Maciej Staszak[1] | Katarzyna Staszak[1] | Karolina Wieszczycka[1] | Anna Bajek[2] | Krzysztof Roszkowski[3] | Bartosz Tylkowski[4,5]

[1]Institute of Technology and Chemical Engineering, Poznan University of Technology, Poznan, Poland

[2]Department of Tissue Engineering, Collegium Medicum, Nicolaus Copernicus University, Bydgoszcz, Poland

[3]Department of Oncology, Collegium Medicum Nicolaus Copernicus University, Bydgoszcz, Poland

[4]Department of Chemical Engineering, University Rovira i Virgili, Tarragona, Spain

[5]Eurecat, Centre Tecnològic de Catalunya, Chemical Technologies Unit, Tarragona, Spain

Correspondence
Bartosz Tylkowski, Eurecat, Centre Tecnològic de Catalunya, Chemical Technologies Unit, Marcel·lí Domingo s/n, Tarragona, 43007, Spain.
Email: bartosz.tylkowski@eurecat.org

Funding information
"Excellence Initiative Research University", project at Nicolaus Copernicus University in Torun (Life Sciences Emerging Field "Multifactorial Molecular-Behavioral Cancer Profiling Team"); Polish Ministry of Science and Higher Education, Grant/Award Numbers: 0912/sbad/2000, 0912/sbad/2010

Edited by: Peter Schreiner, Editor-in-Chief

Abstract

The paper presents a comprehensive overview of the use of artificial intelligence (AI) systems in drug design. Neural networks, which are one of the systems employed in AI, are used to identify chemical structures that can have medical relevance. Successful training of neural networks must be preceded by the acquisition of relevant information about chemical compounds, functional groups, and their possible biological activity. In general, a neural network requires a large set of training data, which must contain information about the chemical structure–biological activity relationship. The data can come from experimental measurements, but can also be generated using appropriate quantum models. In many of the studies presented below, authors showed a significant potential of neural networks to produce generalizations based on even relatively narrow training data. Despite the fact that neural network systems have been known for more than 40 years, it is only recently that they have seen rapid development due to the wider availability of computing power. In recent years, there has been a growing interest in deep learning techniques, bringing network modeling to a new level of abstraction. Deep learning allows combining what seems to be causally distant phenomena and effects, and to associate facts in a way resembling the human mind.

This article is categorized under:
Computer and Information Science > Chemoinformatics

KEYWORDS

artificial intelligence, chemical structure, drug design, machine learning, neural network

1 | INTRODUCTION

Currently, there is an unprecedented range of possibilities for using computers in the medical field—not only as a device to store health records and patient databases or to operate medical equipment but, above all, as a tool to support diagnosis and drug design. This, of course, requires advanced techniques employing artificial intelligence (AI). The advent of the popularity of deep neural learning dates back to 2012 when Krizhevsky et al.[1] won the Large Scale Visual Recognition Challenge.[2] Artificial neural networks providing diagnostic, identification, and organizational potential, especially for large clinical and biological datasets, are becoming increasingly used in medical science. Drug discovery,[3–10] lead optimization[11] and synthesis,[12,13] cardiological and cardiovascular diseases,[14–18] medical image analysis,[19–22] diabetic diseases,[23,24] oncology research,[25,26] diagnosis, for example, alteration of oscillatory brain activity as a possible biomarker for use in Alzheimer's disease diagnosis,[27] are some of the examples of AI in service of medical science (Figure 1). Computer-aided drug design is not only an interesting concept but also a business requirement. As was noted by Wong and Siah,[28] based on a sample of 406,038 entries of clinical trial data for over 21,143 compounds from years 2000–2015, only a small percentage of substances tested are commercially successful and can be used by the pharmaceutical industry. For example, the probability of success (POS) for an orphan drug is 6.2%, and ranges from a minimum of 3.4% for oncology to a maximum of 33.4% for vaccines (infectious diseases). This low success rate encourages the search for alternative ways to design drugs. In several cases, the widely cited statistics present more optimistic values of POS than those by actual databases. For example, in the aforementioned oncological example, authors show that pre-2019 studies had even larger (5.1%) success rate than those from 2019 (3.4%).

Biological systems are a complex and rich source of information, especially in the field of human diseases. Such information has been systematically measured and collected using various technologies to reach an unprecedented volume. The emergence of such high-performance research approaches in the field of biology and diseases creates both challenges and opportunities for the pharmaceutical industry. This is primarily because of the increased possibility of identifying reliable therapeutic hypotheses which can be the basis for developing appropriate drugs. The recent huge growth in computing capabilities has led to an increased interest in machine learning (ML) techniques and their use in the pharmaceutical industry.[29] Creating computer platforms in a distributed architecture gives a virtually unlimited access to storage and a large increase in computing power required for effective learning of complex AI systems. It also enables processing of many types of large data sets constituting the basis for building reliable ML systems. Such data may include text, images, spatial representations of medical scans, biometric data and other information from research and diagnostic work, as well as data on multidimensional elements often represented in tensor form. Much of this rapid increase in performance is due to the high availability of computer hardware, especially graphics processing units (GPUs), which significantly accelerates computing, particularly in the area of parallel processing. The ability of a software tool to use the GPU depends on the source of such application. Closed source software, for example, AlphaFold gives the possibility to use GPU architecture, on the other hand the Polypharmacology Browser 2 (PPB2) technical specification does not contain any data about this possibility. Many of the available applications (Chemputer, DeepChem, DeepNeuralNet-QSAR, DeltaVina, NeuralGraphFingerprints, Open Drug Discovery Toolkit [ODDT], Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry [ORGANIC], REINVENT, SCScore,

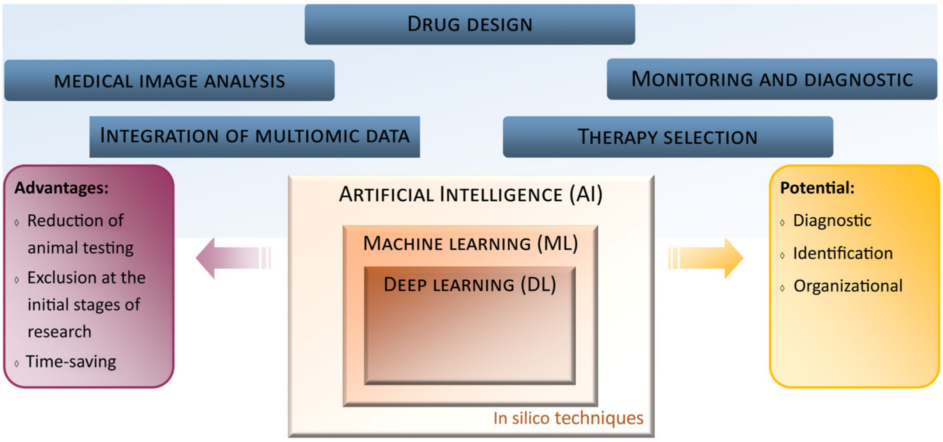

FIGURE 1 Artificial intelligence in medical application

SIEVE-Score, QML) are open-source applications implemented in Python. Therefore, it is up to the user/researcher to decide whether to run the code using an interpreter that performs calculations in the traditional way on the CPU or on the GPU. GPU vendor Nvidia has pioneered the use of GPU-based interpreters to run SciPy or NumPy packages included in drug design tools.

The current rapid development of ML algorithms, such as deep learning (DL), which allows building complex and flexible models based on data, and the success of these techniques in numerous and sometimes very distant areas, have further contributed to a huge increase in interest in ML in pharmaceutical companies over the last few years.[30] In general, AI is the broadest concept, also in health care[31] including problems related to training of neural networks. ML is a slightly narrower area usually considered to encompass algorithms using large, extensive neural networks, but designed for selected problems. Currently, the narrowest classified area of AI systems is DL. The main difference between ML and DL is the level of abstraction of the problems that these techniques cover. For example, an ML algorithm can be used to study the relationship between the structure and a selected physicochemical property of a substance. In turn, a DL algorithm can be used to give an answer on the potential relationship between disease symptoms and the structure of a therapeutically active compound required for treatment, which requires a much more complex neural network to cover a higher level of abstraction. DL algorithms are used to combine causally distant events and effects, a task that has so far only been possible for humans. Although ML and DL are in some aspects similar AI approaches the important distinction between them is the scope and complexity of the problems that they can operate on. ML is a broader term that includes DL in its meaning. DL, however, has capabilities that are broader than ML, due to the use of a larger network structure and its greater complexity. In general, ML gives the ability to create classification models however only with the provision of appropriate features whereas DL gives the ability to generate classification features on its own. DL is used to solve much more complex problems where the datasets are huge, characterized by high diversity and the data is less structured. A significant advantage of DL over ML is that as learning progresses, the network learns to extract features independently eliminating the need for manual feature extraction. However, it should be noted that DL requires significantly more hardware resources. This review focuses on recent developments in the field of drug design based on AI.

2 | STAGES OF DRUG DESIGN

Hughes et al.[32] present the typical stages of drug discovery (Figure 2). Exploration of available biomedical data has significantly intensified target identification. Data mining, refers to the use of a bioinformatic approach to improve identification, and also to indicate and prioritize potential targets for diseases.[33] Another effective method is the search for potential genetic relationships, for example, between genetic polymorphism and the risk of disease or disease development, or establishing whether a polymorphism is functional.[34] A further method is the use of phenotypic screening to determine disease-relevant objectives. Phage display is one of the most potent and extensively used laboratory techniques for studying protein–protein, protein–peptide, and protein–DNA interactions. This approach is mostly based on the protein display on the surface using phages, and is then used to investigate purpose-built libraries containing millions or even billions of bacteriophage that were displayed.[35]

Validation methods include techniques from in vitro approaches through the use of complete animal models and modulation of a required objective in ill patients. Certainty in the observed results is considerably increased by a multi-validation approach. After the target validation process, complex screening tests are developed during the “hit” identification phase and the main discovery in the drug discovery process. A “hit” molecule is defined as a chemical

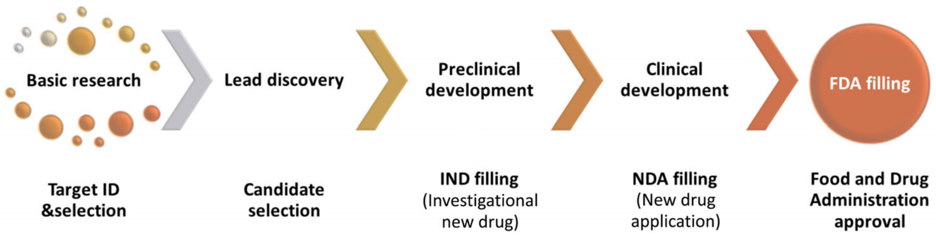

FIGURE 2 Stages of early drug discovery

compound which has the desired activity in a complex screening test and whose activity is proven after reexamination. A method called high throughput screening (HTS) involves screening of the whole chemical component library directly in relation to the drug target. Alternatively, a more complex assay arrangement is used, such as a cell-based assay, in which the activity is target-dependent, but which consequently would also need secondary assays to verify the mechanism of action of the investigated substance.[36]

Besides HTS, compound libraries, such as datasets based on rule of five,[37] are used to determine the “hit” set of molecules for further investigations. Analysis of the substance “hit” list based on algorithms of computational chemistry permits refining and selecting hits for further progression based on a chemical cluster understood as an ensemble of molecules and factors such as ligand performance, which give an idea of how well a compound produces an effect of required or expected magnitude. This is followed by the “hit to lead” phase in which the effort is made to extract more effective and selective compounds from the hit series, such that they exhibit properties sufficient to test their efficiency in any in vivo models available. Normally, this task involves carrying out intensive structure–activity relationship (SAR) studies of each of the main complex structures, with measurements to determine the activity and selectivity of each individual compound. Quantitative structure–activity relationship (QSAR) and SAR models[38–42] are mathematical modeling techniques that can be used to predict physicochemical properties and biological activities for the analyzed chemical compounds based on their known chemical structure. These models are available free of charge or as commercial computer programs. QSAR models must be scientifically valid, and the substance must belong to the field of application of the model. The aim of this final phase of drug discovery is to preserve the promising features and characteristics in lead components while making improvements of flaws in the lead structure. All molecule data collected at this stage will allow the development of the final candidate profile which, along with the toxicological and chemical production and control conditions, will provide the basis for a regulatory application to start administering to humans.

In the case of drug design, AI is used primarily to assess the potential properties of active substances and, to a lesser extent, to discover new drugs or new uses for already existing drugs (drug repurposing) and synthesis routes. At each of these steps it is necessary to know the structure of the compound, and its interactions.[43]

3 | DRUG DESIGN IN PRACTICE

It was shown by Lipinski,[37] who introduced a rule of five which defines molecular properties essential for a drug's pharmacokinetics in the human body, that the chemical space might contain as many as 10^{60} compounds when taking into consideration only basic structural rules.[44] In the light of the above, researchers have been creating databases of drug like chemical structures. The biggest databases are GDP-13,[45] containing approximately 970 million compounds, and GDP-17,[46] containing 166 billion organic small molecules, both freely available for researchers. There are also databases created purely on an ab initio basis using quantum calculations. Maho[47] derived a database containing 1.52 million substances using a density-functional theory (DFT) approach with the B3LYP exchange-correlation functional and basis set 6–31+G* able to represent electronic wave functions of chemical elements up to argon. Such databases create the potential to research possible pathways for drug design.

One of the key problems that occurs when comparing chemical compounds for selected structural features is the relatively high complexity of the process of searching for and identifying selected chemical substructures. It is assumed[48] that searching for chemical structures belongs to the class of non-polynomial-complete computational problems $O(k^N)$, where N is the number of atoms. This means, in a worst-case scenario, an exponential increase in the duration of calculations with each successive atom added to the investigated structure. In the traditional approach, the solution used to describe similarities between substances was to capture the structure in topological indices, for example the Wiener, Balaban or Hosoy index. In strict applications in drug design procedures, topological indices carried too little information about the indexed compound. The solution proposed was then to use structural keys, in which appropriate information about structure was encoded using bit string expressions. The disadvantage of structural keys, however, was the requirement to employ a definite and unique agreement on how to code chemical structures, which limited their level of generalization. To overcome this problem, a higher level of abstraction was proposed in the form of molecular fingerprints, in which the necessity of using predefined patterns was eliminated and which, consequently, enabled the generalization without the use of predefined patterns. Similar to cryptographic fingerprints, a given chemical substructure is represented by a numerical hash being a sequence of bits. The specific binary representation of a given substructure is irrelevant, but it is important that each substructure is represented equally. The coding of the molecular fingerprint is done by means of a specific, typically proposed by researchers or software developers, randomization function, also called the hash function. Hashed fingerprints are a type of black box encoding a structure, which at the same time ensures that similar substructures receive a similar set of bits representing them. For example, convolutional neural networks (CNNs) are used in many areas of medical expertise.[49,50] CNNs can progressively filter different portions of training data and refine important features in the discrimination process used to recognize or classify patterns. A typical artificial neural network using neural connections on any-to-any basis can easily be overtrained. However, in the CNN's convolutional substructure, each neuron is connected only to the local input region. Local areas are defined by width and height, while depth extends through the entire input image layer. Such a limited area of hyper-connections is called a reception area. Convolutions allow extracting simple features in the initial layers of the network, for example, during image processing they recognize edges with different orientation or areas with different colors, and then shapes and geometric objects in the subsequent layers. Convolutional layers perform mathematical convolution operations on the input data and pass their results on to the next layer. This is similar to the reaction of the neuron in the visual cortex to a specific stimulus. Exemplary, processing of data describing a chemical structure involves recognizing its fragments by the convolution layer in individual iterative steps (Figure 3), thereby identifying individual characteristics of the structure. Such a network can be trained by applying input data encoded with SMILES (simplified molecular input line entry specification) notation[51] or even a bitmap image of a classical chemical structure. Eventually, the convolutional network is able to learn the relationship between the structure and the target parameter of biological, chemical, or physicochemical activity.

Various neural network algorithms have been proposed in the literature in drug design, ranging from very simple to extremely complex.[52] Due to the enormous breadth of the topic of neural network algorithms, only a very brief summary of the models discussed in this paper is given below. Some of the simple ones include multilayer perceptron (MLP) based on McCulloch-Pitts neurons or more complex regression classifiers such as logistic, naive Bayes, shallow neural networks, ridge, lasso, or support vector machines (SVM). Logistic regression is used when a variable is dependent on a dichotomous scale[53] and the explanatory variable has a two-point distribution. Naive Bayes regression is linear method in which statistical analysis is carried out using the method of Bayesian inference[54] and classifiers are based on the assumption of mutual independence of independent variables. Shallow network models usually have up to two layers of neurons and require properly prepared features to perform the learning process. Such models are relatively easy to overtrain which is characterized by too faithful adherence to specific data that such a network has already observed. Ridge and lasso regression is a type of regularization that consist on introducing additional information to the ill-conditioned problem to improve the quality of the solution and is used as a method to increase the generalization

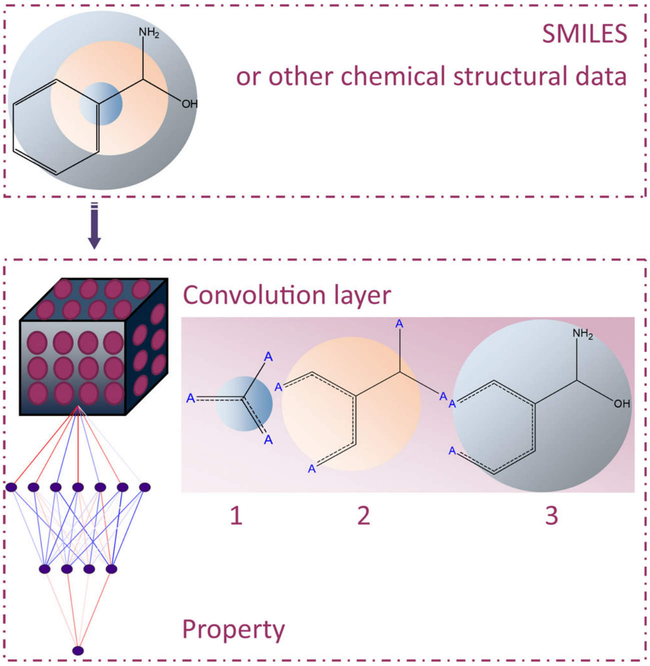

FIGURE 3 Illustration of identification of chemical substructures by the convolutional layer of neural network

of the trained model. Lasso (least absolute shrinkage and selection operator, L1) originally proposed by Santosa et al.[55] is able to reduce variability and improve the quality of linear regression methods. Ridge (Tikhonov regularization, L2) regression[56] is a method used when independent and explanatory variables are strongly correlated. The standard errors of ridge regression are reduced by the addition of a certain amount of bias to regression estimates. Besides SVM is a type of kernel class of algorithms and an abstract concept of a machine that acts as a classifier, whose training is to determine a hyperplane which separates examples belonging to two classes with a maximum margin.[57] Taking into account separate objects, the kernel function evaluates a certain similarity measure. Generalized regression neural network (GRNN) is a neural network that combines the advantages of radial network and MLP. The first hidden layer utilizes radial neurons, performing clustering of the input data. The second layer consists of only two summation neurons and is called the regression layer. Interesting solution is used hierarchical linear models which allow to take into account the structure of relationships between variables grouping observations.

Among the techniques used in drug design, tree based models also show satisfactory results. The decision trees include many learning algorithms to express given hypotheses.[58] Decision trees are widely used in problems concerning classification and prediction of ideas and concepts, among others in medical diagnostics. Random forest[59] involves constructing multiple decision trees while teaching and generating a class which is a mode, indicating the value with the highest probability of occurrence, or the value most frequently occurring in the sample, or predicted average of individual trees.[60] Random forests are a way of averaging many deep decision trees, trained on different parts of the same training set, to reduce variance.

The relatively simple radial networks and their more elaborate successors are a different way of solving learning problems. Radial network is a type of unidirectional neural network in which radial basis functions (RBF) is used and radial neurons are applied. RBF are real functions whose value depends on the distance from a certain point, that is, it is a measure of distance. A representative radial network contains an input layer, a hidden layer consisting of radial neurons and an output layer, working out the network response. Radial neurons are used to recognize repetitive and characteristic features of clusters of input data. More complex models that utilizes radial networks are probabilistic neural networks (PNN) in which the number of neurons in the hidden layer is equal to the number of training cases. The main feature of probabilistic networks is to normalize the values of output signals in such a way that their sum on all outputs of the network has the value of one. It can then be assumed that the values on the individual outputs of the network represent the probabilities of categories assigned to those outputs.

The most complex DL network models include GANs, CNNs, and capsule networks.

GANs are type of networks used to create very realistic content. GAN consists of two parts, a generator and a discriminator, which engage in competition with each other during training. The generator creates artificial content and the discriminator tries to distinguish it from real world data. The network is trained to the point where the discriminator cannot differentiate the artificial data from the real data.

Capsule networks proposed by Geoffrey Hinton[61] improve generalization to new points of view, which means that after training in handling rotation, they learn that an object can be viewed from several different sides. Single computing unit in capsule networks is a capsule which is a generalized type of neuron. Vector carries information about the strength of activation through its length and about the context of activation through its direction.

3.1 | Determining drug properties

The great potential of AI was recognized by the pharmaceutical industry and medical community several years ago. There have even been large programs integrating scientists, which have made it possible to create large databases which form the basis for ML. An excellent example of this is the Tox21 Data Challenge[62] containing details of 12,000 environmental chemicals and drugs, including 12 different toxic effects, comprised stress response effects and nuclear receptor effects. Stress response panel consisted of the nuclear factor (erythroid-derived 2)-like 2 antioxidant responsive element, heat shock factor response element, genotoxicity indicated by ATAD5, mitochondrial membrane potential, and DNA damage p53 pathway. Nuclear receptor panel (biomolecular targets) contained the following elements: estrogen receptor alpha; androgen receptor; estrogen receptor alpha, luciferase; androgen receptor, luciferase; aryl hydrocarbon receptor; peroxisome proliferator-activated receptor gamma; aromatase. Based on 12,000 compounds as training data, ML was proposed for the evaluation of 647 compounds with excellent accuracy. It should be noted that the success of learning methods largely depends on the training datasets which are, therefore, the first prerequisite for obtaining reliable models.

In silico methods in drug properties prediction are based on several techniques. Artificial neural networks are the main method proposed for QSAR models (Figure 4).[15] This technique is widely used by the pharmaceutical industry in the drug discovery process. As early as at the beginning of the century, scientists noted that increasing computer power can support decision making in this area. For example, a study[63] compared a SVM with ML methods (RBF kernel and C5.0 decision tree) in predicting inhibition of dihydrofolate reductase by pyrimidines. The authors showed that SVM is an effective deterministic learning algorithm with reproducible results, with the lowest model error, as well as the shortest calculation time compared with the RBF ML methods. Based on this methodology, it is possible to predict the properties of drugs in the context of their toxicity.

Chemical carcinogenesis prediction is very important in drug discovery because of the crucial impact of drugs on human health.[64] In this case, two main mechanisms are considered: genotoxicity (by the mutagenicity of DNA-damaging chemicals) and non-genotoxic carcinogenic action. Distinction between both mechanisms is very important for risk assessment. It is crucial for non-genotoxic carcinogens which are classified as promoters for tumor development. However, genotoxicity is a risk factor at different concentrations and may result in mutations causing tumor growth initiation. Many recent studies have shown that environmental factors, including various chemicals, play a key role in cancer development.[65] Therefore, it is extremely important to identify substances with such activity and to prevent exposure to such carcinogens. Traditionally, animal assays were used to indicate substances with the carcinogenic potential. However, this method is not only costly and time-consuming, but also complicated by regulatory policies demanding changes in protocols of examination of toxicological effects. Singh et al.[66] showed a possibility to use the PNN and GRNN modeling approaches in prediction of carcinogenicity of diverse chemicals (by determining the tumorigenic dose, $-\log TD_{50}$). Authors employed the dataset from Carcinogenic Potency Database[67] including: for rats, 834 compounds (466 positive and 368 non-positive carcinogens), for mice, 632 (292 positive), for hamsters, 57 (38 positive). Of the various molecular descriptors, 12 non-quantum mechanical molecular descriptors were used. They could be divided into four categories: (i) physicochemical (octanol–water partition coefficient as Log P, density, melting point, half-life in water or in air, persistence time), calculated by molecular structures; (ii) constitutional (hydrogen-bond acceptor or donor, and carbon or hydrogen percentage); (iii) geometrical (maximum Z-length); and (iv) topological (Balaban index), computed based on 2D structures of the molecules (in the form of SMILES). It should be noted here that the authors employed relatively simple descriptors based mainly on physical and chemical properties for the evaluation of complex final estimators, prediction of carcinogenicity. Both models proposed differ in architecture, 5 or 9 input, and hidden layer for PNN and GRNN models, respectively. Moreover, PNNs are based on the Bayesian classification and classical estimators for probability density function,[68] while GRNNs are trained by a K-means clustering algorithm. The authors showed that the optimum PNN exhibited a high ability to predict and differentiate substances between positive and non-positive carcinogens and may be treated as a preliminary stage for the possible exclusion of new substances with a carcinogenic potential. The GRNN model, on the other hand, allowed predicting the tumorigenic dose with high accuracy.

These relatively simple models were presented in a study[66] and initiated further research in this area. For example, various ML models of in vitro and in vivo bioassays for rat carcinogenicity prediction were presented in reference 69. The first advantage over the previously described models is that here the authors used a much larger set of training data, including GreenScreen with genotoxicity results (in vitro GADD-45a-GFP assay) for 1415 compounds,[70] Syrian Hamster Embryonic with in vitro Syrian Hamster Embryonic (pH 7+) cell transformation assay results (356 compounds),[71] Hansen Toxicity Benchmark dataset with Ames bacterial mutagenicity results (6512 compounds),[72]

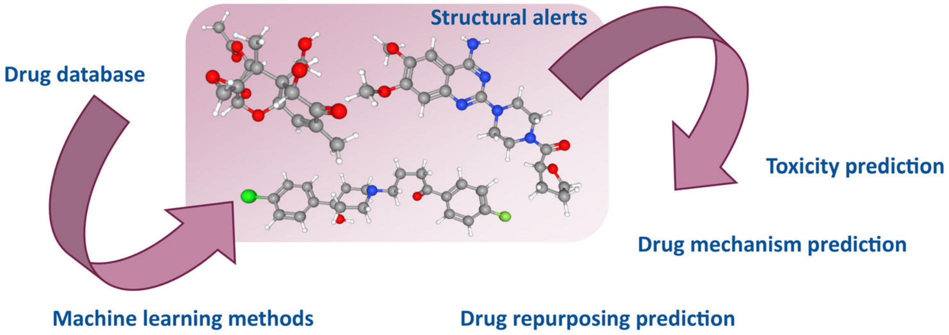

FIGURE 4 Machine learning methods in drug design

ISSCAN (in vivo rat carcinogenicity, 854 compounds),[73] in vivo rodent pharmaceutical carcinogenicity results (374 compounds).[74] Moreover, to predict the assay results in this case, higher numbers of ML algorithms were compared: J.48 Decision Tree, Random Forest, MLP, k-nearest neighbor and Adaboost[75] with 10-fold cross validation. Moreover, descriptors associated with physicochemical properties were used to describe the compounds. The authors used for this purpose, such parameters as: (i) presentations of chemical structures by ChemAxon Standardizer with redrawn 3D coordinates, the explicit representation of hydrogens and reconfigured aromaticity; (ii) physicochemical properties (octanol–water partition coefficient as Log P), number of hydrogen-bond acceptor or donor, as well as rotatable bonds, polarizability, polar surface area, and molecular weight. It should be noted that the assessment of properties of substances (potential drugs), that is, in the context of carcinogenicity, is always supported by the assessment of chemical properties. It should be noted that the assessment of properties of substances (potential drugs), that is, in the context of carcinogenicity, is always supported by the assessment of chemical properties. This confirms the important role of coordination of chemistry in drug design with the use of ML techniques. The authors concluded that k-nearest neighbors model was the best one of all considered for in vivo rodent carcinogenicity prediction, and that the results obtained can contribute to future development of new drugs and determination of their properties with AI methods.

From the point of view of drug design, acute toxicity analysis is important as well. This parameter indicates unequivocally whether it is worth considering a given substance as a potential drug or whether, in view of the strong hazard to human health, any further stages of research in this area should be abandoned. It could also predict the side-effects of overdosage and should support all phase III clinical trials of drugs. Evaluation of acute toxicity could help in the identification of patients at higher risk for overdosing, for example, those suffering from depression or dementia. In terms of acute oral, dermal and inhalation toxicity, the most common studies reported in the literature are related to oral toxicity assessment (mainly as median lethal dose, LD_{50} parameter). In these studies, authors mainly use the database created by Zhu et al.[76] for 7385 compounds with their most conservative lethal dose. For example, in one study, authors predicted oral acute toxicity based on a molecular graph encoding a convolutional neural networks standard (MGE-CNN) with regression model, a multi-classification model and a multi-task model for deep fingerprints.[77] Analysis of data allowed extracting structural fragments of molecules responsible for toxicity: nitriles, alyl (thio)phosphates and thicarbonyl. The presented DL architecture for acute oral toxicity could be used for prediction and exploration of other toxicity or property endpoints of chemical compounds. Moreover, by using the ability to learn automatically from DL, it was also possible to create fragments from information about atoms and bonds and then identify their potential toxicity. Researchers from Peking University, Center of Quantitative Biology and Molecular Design Laboratory, made these DL models available at a website.[78] The issue of oral toxicity prediction was also discussed in.[79] The authors showed the superiority of dual-layer hierarchical models (by integration regression and classification QSAR models) over classical base models in the prediction of categories (binary toxic/nontoxic and four hazard categories under the U.S. Environmental Protection Agency [EPA] classification system) and continuous (LD_{50}) endpoints for rat acute oral toxicity. The first layer of the proposed model was based on regression, binary and multiclass ML techniques, and molecular descriptors and fingerprints, while the second one was based on collection of the outputs from the base models. In order to confirm the validity of the adopted learning model, the authors presented calculations for two substances: Furaserenon-X ([(1*S*,2*R*,3*S*,7*R*,9*R*,10*R*,11*S*,12*S*)-3,10-dihydroxy-2-(hydroxymethyl)-1,5-dimethyl-4-oxospiro[8-oxatricyclo[7.2.1.$0^{2,7}$]dodec-5-ene-12,2′-oxirane]-11-yl]acetate) and VX (Ethyl({2-bis(propan-2-yl)amino]ethyl}sulfanyl)(methyl)phosphinate). Furaserenon-X is a class of trichothecene mycotoxins. It causes disruption of DNA synthesis by inhibiting protein synthesis,[80] with logLD_{50} at the level of -1.95 mmol/kg (EPA, class I). The base regression model predicted that this compound is nontoxic (EPA, class III, log$\text{LD}_{50} = -0.39$ mmol/kg), while the hierarchical classification model identified it as toxic (EPA, class I, log$\text{LD}_{50} = -1.14$ mmol/kg). VX is an extremely toxic class of organophosphorus compounds belonging to thiophosphonates (EPA, class I, log$\text{LD}_{50} = -4.34$ mmol/kg), which potentially blocks the function of acetylcholinesterase. As a consequence, flaccid paralysis of all muscles in the body occurs. The immediate cause of death is asphyxiation caused by paralysis of the diaphragm muscle.[81] In this case, both tested and comparable models were not very accurate, although both correctly predicted toxicity (EPA, class I). It should be noted that the low accuracy of prediction was due to the small amount of training data in such a high toxicity range. However, a smaller error could also be seen here for the hierarchical model (log$\text{LD}_{50} = -1.19$ mmol/kg) in comparison with the base model (log$\text{LD}_{50} = -1.08$ mmol/kg). It is worth pointing out that the artificial neural network should have the ability to generalize, just like humans, and the low accuracy of prediction obtained indicates that the training data did not include the range of testing data. Thus, the main problem here is the volume of database. Alberga et al.[82] proposed prediction of toxicology endpoints related to the acute oral systemic toxicity as binary classification: nontoxic ($\text{LD}_{50} > 2000$ mg/kg) and very toxic ($\text{LD}_{50} < 50$ mg/kg), as well as classification according to EPA and GHS

(Globally Harmonized System of Classification and Labeling of Chemicals) based on k-nearest neighbors techniques and 19 different fingerprints (Table 1). The authors concluded that the increasingly accurate methods of predicting acute oral toxicity may replace the necessary animal tests.

Cardiotoxicity is often described with regard to blockade of human ether-à-go-go-related gene (hERG) cardiac potassium channel.[89] Cardiovascular toxicity comprises heart failure due to toxin-induced abnormalities with injury of the muscles, and therefore may reduce blood flow and circulation. It is also the main reason for withdrawal of many drugs from markets globally. Lengthening of the QT interval related with lethal ventricular arrhythmia is responsible for such situations. Since this aspect is very important in drug design (mainly as safety evaluation of drug candidates), in silico methods are described in the literature.[90] Zhang et al.[84] proposed prediction of hERG activity by deep neural networks (optimal form of calculation with three hidden layers) based on 697 molecules data from.[85,91] Based on the results, the authors concluded that the proposed DL could offer effective prediction of hERG toxicity, and as a consequence, have a great potential to aid developing novel drug candidates. Similar observations were described in another paper[86] looking at ML and DL algorithms using fingerprints and principal component analysis (including partition coefficient, molecular weight, H bond acceptors and donors, number of rotatable bond, rings and aromatic rings, as well as molecular fractional polar surface area) as descriptors and a training set of 3991 compounds. Authors compared SVM methods (linear, polynomial, radial), with random forest, and artificial neural network (layer size 100, 200, and 400) for DL. Based on the results, it can be seen that accuracy of hERG-blocker prediction depends on the selection of

TABLE 1 Targets, selected descriptors, and statistics (classification accuracy, sensitivity, specificity) for selected models

Target	Descriptors	Statistics	Ref.
Carcinogenicity prediction[a]	H bond acceptors H bond donors Content of H and C	Sensitivity 89.6%, specificity 95.8%, accuracy 92.09%	66
	H bond acceptor H bond donors Rotatable bonds Polarizability Polar surface area	Sensitivity 35.1%, specificity 88.3%, accuracy 69.3%	69
Oral acute toxicity	Molecular fingerprint	Accuracy 95.5%	77
		Accuracy 71%	79
	Molecular fingerprint (e.g., atom pairs, topological torsion, substructure, hybridization)	Sensitivity 83.9%, specificity 99.6%, accuracy 82%	82
Cardiotoxicity	Molecular fingerprint and 2D ChemoPy[83] descriptors (e.g., connectivity, topology, Kappa, Burden) or MOE 2D descriptors (surface areas, connectivity, shape indices, atom, and bond counts)	Accuracy 78%	84
	Molecular fingerprint	Sensitivity 78%, specificity 61%	85
	Molecular fingerprint and principal component analysis (PCA) (e.g., H bond acceptors, H bond donors, rotatable bond number, number of rings and aromatic rings, molecular fractional polar surface area)	Accuracy 87%	86
	Molecular descriptor Molecular fingerprint Molecular graph-based features (atom types, number of degrees, number of bound hydrogens, implicit valence, size of ring containing the atom, and aromaticity)	Sensitivity 83.3%, accuracy 77.3%	87
	SMILES and molecular fingerprint (number of tertiary amines (aliphatic), Wiener index, number of carbon atoms, frequency of C–C at topological distance, distance/detour ring index, centered Broto–Moreau autocorrelation weighted by van der Waals)	Accuracy 90.1%	88

Note: Where Sensitivity = TPTP+FN·100%, specificity = TNTN+FP·100%, accuracy = TP+TNTP+TN+FP+FN·100%, TP and TN are the number of true positives and negatives, respectively, while FP and FN are the number of false positives and negatives, respectively.

aCarcinogenicity prediction, as tumorigenic dose (TD50) in reference 66 and in vivo rodent carcinogenicity (IVRC) in reference 69.

fingerprints. Better results for ML models were obtained with the use of integer-type fingerprints, while binary-type fingerprints are appropriate for DL. Ryu et al.[87] proposed a step further—model that predicts both hERG-blockers and non-blockers for input compounds (DeepHIT). The criterion indicating the blocking or non-blocking properties of hERG was the value of the half maximal inhibitory concentration (IC_{50}): hERG-blockers had $IC_{50} < 10$ μM, hERG non-blockers had $IC_{50} \geq 10$ μM.[92] The calculations required a preliminary standardization of the compounds by selection of the largest fragment, removal of explicit hydrogens, ionization, and calculation of stereochemistry. The authors compared six traditional ML algorithms (i.e., k-nearest neighbors, logistic regression, naive Bayes, shallow neural network [simpler configuration, less neural layers], random forest, and SVM) with deep multilayered neural network with molecular descriptor-, molecular fingerprint-, and graph-based feature datasets (Table 1). Their proposals are available at websites.[93,94] In the case of this cited study, it is worth emphasizing that the authors, based on the trained network, indicated a new novel urotensin II receptor antagonists without hERG-blocking activity obtained from a seed compound of a previously reported UT antagonist (KR-36676) with a strong hERG-blocking activity. Capsule networks also showed excellent performance in the classification of hERG-blockers and non-blockers with prediction accuracies of approximately 92%.[95] This is the first example of using such a technique in drug discovery-related studies. Furthermore, work[88] presented an interesting comparison of ML prediction (linear regression, ridge regression, logistic regression, naïve Bayes, neural network, and random forest) for results regarding 10 drug compounds (Table 2). The presented model correctly predicted 8 out of 10 compounds with 80% accuracy, 60% sensitivity and 100% specificity, which indicated that this model could be used for virtual screening in drug discovery.

It should be noted that in all cases of these cardiotoxicity predictions, the chemical structure of the compounds played an important role. The analysis indicates that most hERG channel blockers have in their structure a tertiary amine group and aromatic rings. The first fragment has the ability to protonate at physiological pH and plays a significant role in the binding of the channel blocker and the hERG channel. Aromatic rings are associated with π-stacking or hydrophobic interactions with the aromatic rings of amino acids within the hERG channel cavity.[96]

Known for their ability to be creative, generative adversarial neural (GAN) networks have also found application in de novo drug design. Based on compound databases such as ChEMBL or ZINC Database, the application of these networks allows the generation of new structures—drug-like compounds which can be treated as potential new drugs with desired properties.[52,97,98] For example drug-like Prykhodko et al.[99] successfully proposed latent vector based generative adversarial network (LatentGAN), combination of autoencoder and Wasserstein GAN, for generation of drug-like compounds (set limited to SMILES of containing only [H, C, N, O, S, Cl, Br] atoms and a total of 50 heavy atoms or less) and target-biased compounds (EGFR, HTR1A and S1PR1 targets, based on ExCAPE-DB). The authors of this paper indicated that the proposed model allows the prediction of compounds according to the planned target, while also indicating that a significant portion of the compounds are new with respect to the training set. Another example of adapting the GAN for drug design is its connection with reinforcement learning (RL), known as Objective-Reinforced Generative Adversarial Networks (ORGAN)[100] or its implementation for inverse-design chemistry (ORGANIC).[101] Based on two drug-likeness indicators: chemical beauty[102] and Lipinski's rule-of-five[37] Aspuru-Guzik group[101] showed that ORGANIC allows to generate molecules (based on SMILES sequences format) which are consistent with a comparable list of FDA-approved drugs in the amount of 148 and 207, for both indicators respectively. Among the substances proposed by the model were very well-known compounds, for example, paracetamol and salicylic acid. There are also

TABLE 2 Prediction results of 10 drug compounds[88]

Drug	In vivo results	Model results
Haloperidol	Toxic	Toxic
Chloropromazine	Toxic	Toxic
Disopyramide	Toxic	Toxic
Cimetidine	Nontoxic	Nontoxic
Terazosin	Nontoxic	Nontoxic
Spironolactone	Nontoxic	Nontoxic
Cefazoline	Nontoxic	Nontoxic
Loratadine	Nontoxic	Nontoxic
Sotalol	Toxic	Nontoxic

proposals in the literature for drug design software based on neural networks, including GANs. For example, MolAICal software can be successfully used to generate 3D structural ligands in the 3D pocket of protein targets.[103] The software is based on two modules: fragments of FDA-approved drugs or from the ZINC drug database are used to train the WGAN model, and then the generated fragments are used to grow 3D ligands in the protein pocket. In this approach molecular docking is used for check the affinities between the generated molecules and proteins. It is worth noting that the software supplies the filter rules, for example Lipinski's rule-of-five, synthetic accessibility (SA) and pan-assay interference compounds (PAINS). Moreover, other user-defined rules can be added. The authors indicate that the proposed software can create ligands with 3D structural similarity to the crystalline ligand of GCGR or SARS-CoV-2 M^{pro} and could become a useful tool for drug design.

Although in silico techniques are still relatively new, they are becoming increasingly important in drug design. On the one hand, the enable the reduction of animal experiments, which is in line with general scientific trends. On the other hand, these techniques enable an initial assessment of the broadly defined toxicity of a compound before it is synthesized, thus at a very early stage of drug design. The above examples of use of such computational techniques for analyzing quantitative structure–activity relationship, ML and DL, undoubtedly justify the use of these methods in the determination of toxicity in silico, especially when there are no experimental results, and the possibility of AI generalization of data is very helpful.

3.2 | Drug mechanism

Another aspect to consider in drug design procedures is predicting the interaction between the drug and the target (enzymes [E], ion channels [IC], nuclear receptors [NR], G protein-coupled receptors [GPCRs], known as gold standard according to Yamanishi et al.).[104] At the same time, such a procedure may facilitate understanding of the drug mechanism of action, pathology of the disease and possible side effects of the drug.[105] In a simplified form, it can be said that the drug binds to the target molecule by formation of temporary bonds and reacts with the target to inhibit its functioning and to avoid certain catalyzed reactions occurring in the body in order to treat diseases. Depending on the type of drug, its molecule interacts directly with the active site of the target to inhibit reaction (competitive inhibitors) or with an allosteric site on the target to change the reaction (allosteric inhibitors).[106] Regardless of the mechanism, assessment of the drug–target interaction (DTI) potential should take into account the structure of both the drug and the target under consideration, together with the possibility of bond formation and reaction.[107] Identification of DTIs is a crucial step in drug discovery. AI techniques are often proposed for the prediction of DTIs thanks to, as mentioned above, the opportunity to use increasingly large databases and the ability of neural networks to generalize.[108] For example, Rayhan et al.[109] proposed connection of two deep, CNNs for DRI prediction: FRnet-Encode and FRnet-Predict. The first of them was used to generate 4096 features for dataset, and the second one to classify and identify probability of interaction with an accuracy of over 97%. This approach of analyzing data using two models is very effective. The authors also proposed new pairs of compounds with a high probability of interaction in all four gold standard datasets (i.e., [E] protein ID hsa:10825/drug: threonine with score: 0.8351, [IC] protein ID hsa: 285242/drug: diazoxide with score: 0.9823, [NR] protein ID hsa: 2099/drug: tazarotene with score: 0.9912, [GPCR] protein ID hsa:9052/drug: isoetharine with score: 0.9013). However, these results were not verified by the authors and are only a hypothesis. An interesting approach was also presented in a study.[110] The authors used DL with convolution on protein sequences to predict DTI. The model proposed by them employed raw protein sequences both for different target protein classes as well as protein lengths. The model, similarly to other cited works, was validated by prediction DTIs from bioassays such as PubChem BioAssays and KinaseSARfari with a high accuracy. Pliakos and Vens[111] presented heterogeneous networks with biclustering trees. They used descriptors based on chemical structure for drugs and descriptors based on the alignment of protein sequences for proteins. This is the most frequently used method, and differences between authors are mainly related to the amount of data of learners, databases, and network topology. As was suggested by the authors, use of tree-ensemble learning models with output space reconstruction allowed obtaining higher prediction results in comparison with traditional models. Moreover, such a solution is known for its scalability, interpretability and inductive setting, which is very important in prediction. Li et al.[112] showed usefulness of combinations of position-specific scoring matrix (including protein secondary structure, protein binding site and prediction of disordered regions) and local phase quantization methods, as well as rotation forest classifier in the prediction of DTIs with average accuracies equal to 89.15%, 86.01%, 71.67%, 82.20% for the four targets, respectively. To confirm the validity of the predictive model developed, the authors tested the algorithm on a commercial drug sulfasalazine (2-hydroxy-5-[(E)-2-{4-[(pyridin-2-yl) sulfamoyl]phenyl}diazen-1-yl]benzoic acid) and two target protein sequences: arachidonate 12-lipoxygenase, 12S-type (ALLOX 12) and lipoprotein lipase (LPL). According to the prediction results, sulfasalazine interacts with ALLOX 12 with a possibility score of 0.844, and does not interact with LPL (possibility score = 0.3200). Another approach is the method named DTiGEMS+ based on graph embedding, graph mining, and similarity-based techniques.[113] Authors proposed a heterogeneous network by connecting the known DTI graph with two complementary graphs based on drug–drug and target–target similarities. An interesting approach was to validate the model on unknown data for a group of drugs and targets (enzymes [E], ion channels [IC], nuclear receptors [NR], GPCRs) and to assess their interactions on the basis of experimental data (e.g., from PubMed identifier, PMID or drugs base, DB) not included in the teaching databases (Table 3). The authors did not mention which descriptors were used in the model. It can be assumed that they were similar to those described in their previous work.[114] In this work, random forest model using heterogeneous graph, containing known DTIs with multiple similarities between drugs and multiple similarities between target proteins, was proposed. Chemical structure fingerprints, Gaussian interaction and side-effect profiles were considered as descriptors for drugs, while amino acid sequence profiles of proteins, parameterizations of the mismatch and the spectrum kernels, proximity within the protein–protein interaction network and the Gaussian interaction profile were descriptors for target proteins. The authors also verified the correctness of the model proposed by analyzing new drugs, with 22 out of 25 being correctly identified. Additionally to the models based on heterogeneous graph, a new DL model multimodal deep autoencoder with a similarity network with drugs as nodes and drug–drug similarity values as the weights of edges was proposed by Wang et al.[115]

The presented literature review indicates the possibility of predicting interaction between the drug and the target, which allows a conclusion that AI, just as in the prediction of drug properties, is an interesting alternative in the process of drug design, as well as drug ranking. In this second case, Geres et al.[116] demonstrated the applicability of ML, based on proteomics and phosphoproteomics data derived from 48 cell lines, for predicting therapy in cancer treatment, by evaluating of >400 drugs for their antiproliferative efficacy in tumor cells.

3.3 | Drug repurposing

Drug repositioning allows finding new uses of existing drugs.[117–119] It is also considered as a suitable method for finding drugs for orphan and rare diseases. This procedure reduces the time needed to place a new drug on the market, simultaneously reducing the time and risk of failure because preclinical development and optimization issues can be omitted in a large scale. There are three stages of the drug repurposing strategy: (i) identification of potential molecule; (ii) preclinical tests—mechanistic assessment of the drug effect; (iii) phase II clinical trials—evaluation of efficacy.[118] In the first of these steps, computer-based calculations can be used successfully. The above-discussed DTI prediction is also beneficial for searching novel uses of existing drugs. Based on the increasing access to medical databases,[120–125] it is possible to analyze drugs in the context of their new uses by means of, that is ligand-based approaches. Another excellent review concerning drug databases which could be helpful in DTI with their advantages and disadvantages has been published.[126] The basis of this method is the assumption that similar compounds have similar biological properties,

TABLE 3 Validation of DTiGEMS+[113]

Drug	Target name	Evidence (PMID or DB number)
Nifedipine	E: CYP2C9 (Cytochrome P450 Family 2 Subfamily C Member 9)	9929518
Metyrapone	E: CYP1A1 (Cytochrome P450 Family 1 Subfamily A Member)	9512490
Nicotine	IC: CHRNA4 (Cholinergic Receptor Nicotinic Alpha 4 Subunit)	17590520 DB00184
Nimodipine	IC: CACNA1S (Calcium Voltage-Gated Channel Subunit Alpha1S)	DB00393
Norethindrone	NR: ESR1 (Estrogen Receptor Alpha)	27245768
Testosterone	NR: PGR (Progesterone Receptor)	23229004 23933754
Clozapine	GPCR: DRD3 (Dopamine Receptor D3)	DB00363
Clonidine hydrochloride	GPCR: ADRA1B (Adrenergic Receptor alpha-1B)	DB00575

thus, in the case of drug design, it could be concluded that similar ligands have similar activities in respect to similar targets.[127] The importance of this issue is underlined by the increasing volume of publications in this area. Patrick et al.[128] proposed a word-embedding-based ML approach for drug repurposing for nine cutaneous diseases (including psoriasis, atopic dermatitis, and alopecia areata) and eight other immune-mediated diseases. Based on the validation results, authors concluded that model could predict new drugs for psoriasis, with the highest prediction scores for budesonide (a corticosteroid, currently used to treat asthma and inflammatory bowel disease) and hydroxychloroquine (an antimalarial drug that is also used to treat lupus and rheumatoid arthritis). Anderson et al.[129] used Bayesian ML models for drug repurposing in chordoma. Of the available data, the mTOR inhibitor AZD2014 was indicated as the most potent against chordoma cell lines (IC_{50} 0.35 μM U-CH1 and 0.61 μM U-CH2). Moreover, two currently FDA-approved drugs, afatinib and palbociclib (EGFR and CDK4/6 inhibitors, respectively) demonstrated synergy in vitro ($CI_{50} = 0.43$), alongside AZD2014 and afatanib which also showed synergy ($CI_{50} = 0.41$) against chordoma cells in vitro. As shown in the paper,[130] ML could be applied for prediction of a new therapeutic system for drugs. The authors proposed a model based on decision trees (Bayesian tree-structured) with several molecular features as descriptors. They showed that both 2D and 3D chemical similarity should be used. It is worth emphasizing here that often authors use only one type of molecular similarity, but according to the conclusions in the cited work, it is better to use both at the same time, because each of them transfers notation of different components to the model. Moreover, the authors used new types of features: drug–gene phenotype similarity and the gene–gene expression profile similarity across different tissues. Such a solution allowed them to obtain 78% precision in the case of top 50 predictions, and 48.2% for 500 predictions. Using the constructed neural network, the authors concluded that the antipsychotic drug fluphenazine is a highly probable drug targeting the PRKDC gene which is a potential target for treatment of ATM-deficient cancer.[131] Thus, fluphenazine used for schizophrenia treatment could be considered in cancer treatment. ML and DL approaches for cancer drug repurposing are also discussed in detail in another paper.[132]

It is worth noting that cases discussed here do not exhaust the number of uses of AI in the analysis of drugs. For example, there is the aspect related to drug metabolite prediction.[133] Any additional information about a drug undoubtedly provides a basis for even better understanding of the potential ingredients of therapeutics and provides excellent support for decisions regarding continuation of research into the development of new active compounds for medications.

4 | CONCLUSION

In 2019, an article in Nature Machine Intelligence was published in which the author indicated that “Now AI is back — this time, apparently, for good.”[134] The review presented in this article confirms this hypothesis. Progress in computers and computational algorithms has become an opportunity to support medicine. In the area of drug design, the widest applications, as indicated in the literature, are found not only in networks with a very basic architecture such as MLP or RBF, but also and especially in networks with a very complex design such as CNN, capsule or GAN. There is no definite indication which network is the best tool for such design purposes. However, DL solutions are currently the most popular, as they are becoming more and more faithful reflection of complex ways of thinking characterized by human mind. DL allows not only to analyze data, but also finds and determines the characteristics of the observed sets on its own, while becoming an increasingly versatile tool to support the course of drug design. And although the role of a computer cannot be overestimated, in the end there is always a human who makes the final decision. The promise made by AI for the future are better drugs, discovered and delivered faster. It should also be noted that in the case of drug design, basic properties of the molecules, for example, bonding, quantum and physicochemical properties, are not the only aspects to be taken into account. Medicines may have multiple biological targets and effects, and their efficiency depend on several factors such as bioavailability, effect of formulation and administration, as well as individual genetic profiles of patients.

ACKNOWLEDGMENT

This work was supported by Polish Ministry of Science and Higher Education under the grants 0912/sbad/2000 and 0912/sbad/2010 and “Excellence Initiative Research University” project at Nicolaus Copernicus University in Torun (Life Sciences Emerging Field “Multifactorial Molecular-Behavioral Cancer Profiling Team”).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Maciej Staszak: Writing - original draft (equal). **Katarzyna Staszak:** Writing - original draft (equal). **Karolina Wieszczycka:** Writing - original draft (equal). **Anna Bajek:** Writing - original draft (equal). **Krzysztof Roszkowski:** Writing - original draft (equal). **Bartosz Tylkowski:** Conceptualization (equal); writing - original draft (lead).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Bartosz Tylkowski https://orcid.org/0000-0002-4163-0178

RELATED WIREs ARTICLES

Machine-learning scoring functions for structure-based drug lead optimization.

In silico toxicology: From structure-activity relationships towards deep learning and adverse outcome pathways.

Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases

REFERENCES

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90. https://doi.org/10.1145/3065386
2. Berg A, Deng J, Fei-Fei L. ImageNet Large Scale Visual Recognition Competition (ILSVRC); 2010.
3. Schaller D, Šribar D, Noonan T, Deng L, Nguyen TN, Pach S, et al. Next generation 3D pharmacophore modeling. WIREs Comput Mol Sci. 2020;10:1–20. https://doi.org/10.1002/wcms.1468
4. Peña-Guerrero J, Nguewa PA, García-Sosa AT. Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. WIREs Comput Mol Sci. 2021;10:1–25. https://doi.org/10.1002/wcms.1513
5. Recanatini M, Cabrelle C. Drug research meets network science: where are we? J Med Chem. 2020;63:8653–66. https://doi.org/10.1021/acs.jmedchem.9b01989
6. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18:435–41. https://doi.org/10.1038/s41563-019-0338-z
7. Chen J, Schmucker L, Visco D. Pharmaceutical machine learning: virtual high-throughput screens identifying promising and economical small molecule inhibitors of complement factor C1s. Biomolecules. 2018;8:24. https://doi.org/10.3390/biom8020024
8. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. Mol Inform. 2016;35:3–14. https://doi.org/10.1002/minf.201501008
9. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today. 2015;20:318–31. https://doi.org/10.1016/j.drudis.2014.10.012
10. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci. 2018;4:120–31. https://doi.org/10.1021/acscentsci.7b00512
11. Li H, Sze K, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. WIREs Comput Mol Sci. 2020;10:1–20. https://doi.org/10.1002/wcms.1465
12. Struble TJ, Alvarez JC, Brown SP, Chytil M, Cisar J, Desjarlais RL, et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. J Med Chem. 2020;63:8667–82. https://doi.org/10.1021/acs.jmedchem.9b02120
13. St. Denis JD, Hall RJ, Murray CW, Heightman TD, Rees DC. Fragment-based drug discovery: opportunities for organic synthesis. RSC Med Chem. 2021;12:321–329. https://doi.org/10.1039/D0MD00375A
14. De Marvao A, Dawes TJW, Howard JP, O'Regan DP. Artificial intelligence and the cardiologist: what you need to know for 2020. Heart. 2020;106:399–400. https://doi.org/10.1136/heartjnl-2019-316033
15. Hemmerich J, Ecker GF. In silico toxicology: from structure–activity relationships towards deep learning and adverse outcome pathways. WIREs Comput Mol Sci. 2020;10:1–23. https://doi.org/10.1002/wcms.1475
16. Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, Yarifard AA. Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. Comput Methods Programs Biomed. 2017;141:19–26. https://doi.org/10.1016/j.cmpb.2017.01.004
17. Pławiak P, Acharya UR. Novel deep genetic ensemble of classifiers for arrhythmia detection using ECG signals. Neural Comput Appl. 2020;32:11137–61. https://doi.org/10.1007/s00521-018-03980-2
18. Slomka PJ, Dey D, Sitek A, Motwani M, Berman DS, Germano G. Cardiac imaging: working towards fully-automated machine analysis & interpretation. Expert Rev Med Devices. 2017;14:197–212. https://doi.org/10.1080/17434440.2017.1300057
19. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005

20. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging. 2016;35:1299–312. https://doi.org/10.1109/TMI.2016.2535302
21. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442
22. Pierrard R, Poli JP, Hudelot C. Spatial relation learning for explainable image classification and annotation in critical applications. Artif Intell. 2021;292:103434. https://doi.org/10.1016/j.artint.2020.103434
23. Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA J Am Med Assoc. 2017;318:2211–23. https://doi.org/10.1001/jama.2017.18152
24. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16. https://doi.org/10.1016/j.csbj.2016.12.005
25. Shipp MA, Ross KN, Tamayo P, Weng AP, Aguiar RCT, Gaasenbeek M, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8:68–74. https://doi.org/10.1038/nm0102-68
26. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA J Am Med Assoc. 2017;318:2199–210. https://doi.org/10.1001/jama.2017.14585
27. Furutani N, Nariya Y, Takahashi T, Noto S, Yang AC, Hirosawa T, et al. Decomposed temporal complexity analysis of neural oscillations and machine learning applied to Alzheimer's disease diagnosis. Front Psych. 2020;11:531801. https://doi.org/10.3389/fpsyt.2020.531801
28. Heem Wong C, Wei Siah K, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20:273–86. https://doi.org/10.1093/biostatistics/kxx069
29. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, et al. Transfer learning for drug discovery. J Med Chem. 2020;63:8683–94. https://doi.org/10.1021/acs.jmedchem.9b02147
30. Shen C, Ding J, Wang Z, Cao D, Ding X, Hou T. From machine learning to deep learning: advances in scoring functions for protein–ligand docking. WIREs Comput Mol Sci. 2020;10:1–23. https://doi.org/10.1002/wcms.1429
31. Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? J Arthroplasty. 2018;33:2358–61. https://doi.org/10.1016/j.arth.2018.02.067
32. Hughes JP, Rees SS, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011;162:1239–49. https://doi.org/10.1111/j.1476-5381.2010.01127.x
33. Réda C, Kaufmann E, Delahaye-Duriez A. Machine learning applications in drug development. Comput Struct Biotechnol J. 2020;18:241–52. https://doi.org/10.1016/j.csbj.2019.12.006
34. Bertram L, Tanzi RE. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. Nat Rev Neurosci. 2008;9:768–78. https://doi.org/10.1038/nrn2494
35. Kurosawa G, Akahori Y, Morita M, Sumitomo M, Sato N, Muramatsu C, et al. Comprehensive screening for antigens overexpressed on carcinomas via isolation of human mAbs that may be therapeutic. Proc Natl Acad Sci U S A. 2008;105:7287–92. https://doi.org/10.1073/pnas.0712202105
36. Fox S, Farr-Jones S, Sopchak L, Boggs A, Nicely HW, Khoury R, et al. High-throughput screening: update on practices and success. J Biomol Screen. 2006;11:864–9. https://doi.org/10.1177/1087057106292473
37. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol. 2004;1:337–41. https://doi.org/10.1016/j.ddtec.2004.11.007
38. Suay-Garcia B, Bueso-Bordils JI, Falcó A, Pérez-Gracia MT, Antón-Fos G, Alemán-López P. Quantitative structure–activity relationship methods in the discovery and development of antibacterials. WIREs Comput Mol Sci. 2020;10:1–13. https://doi.org/10.1002/wcms.1472
39. Davis AM. Quantitative structure–activity relationships. Comprehensive medicinal chemistry III. Volume 3–8. Oxford: Elsevier; 2017. p. 379–92. https://doi.org/10.1016/B978-0-12-409547-2.12348-0
40. Griffen EJ, Dossetter AG, Leach AG. Chemists: AI is here; unite to get the benefits. J Med Chem. 2020;63:8695–704. https://doi.org/10.1021/acs.jmedchem.0c00163
41. Chuang KV, Gunsalus LM, Keiser MJ. Learning molecular representations for medicinal chemistry. J Med Chem. 2020;63:8705–22. https://doi.org/10.1021/acs.jmedchem.0c00385
42. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. J Med Chem. 2020;63:8761–77. https://doi.org/10.1021/acs.jmedchem.9b01101
43. Iglesias J, Saen-oon S, Soliva R, Guallar V. Computational structure-based drug design: predicting target flexibility. WIREs Comput Mol Sci. 2018;8:e1367. https://doi.org/10.1002/wcms.1367
44. Liu R, Li X, Lam KS. Combinatorial chemistry in drug discovery. Curr Opin Chem Biol. 2017;38:117–26. https://doi.org/10.1016/j.cbpa.2017.03.017
45. Blum LC, Reymond JL. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc. 2009;131:8732–3. https://doi.org/10.1021/ja902302h
46. Ruddigkeit L, Van Deursen R, Blum LC, Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model. 2012;52:2864–75. https://doi.org/10.1021/ci300415d
47. Nakata M. The PubChemQC project: a large chemical database from the first principle calculations. AIP conference proceedings. Volume 1702. Melville, NY: American Institute of Physics; 2015. p. 090058. https://doi.org/10.1063/1.4938866

48. Conte D, Foggia P, Sansone C, Vento M. Thirty years of graph matching in pattern recognition. Int J Pattern Recognit Artif Intell. 2004;18:265–98. https://doi.org/10.1142/S0218001404003228
49. Rathi PC, Ludlow RF, Verdonk ML. Practical high-quality electrostatic potential surfaces for drug discovery using a graph-convolutional deep neural network. J Med Chem. 2020;63:8778–90. https://doi.org/10.1021/acs.jmedchem.9b01129
50. Stecula A, Hussain MS, Viola RE. Discovery of novel inhibitors of a critical brain enzyme using a homology model and a deep convolutional neural network. J Med Chem. 2020;63:8867–75. https://doi.org/10.1021/acs.jmedchem.0c00473
51. Weininger D. SMILES, a chemical language and information system: 1: Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28:31–6. https://doi.org/10.1021/ci00057a005
52. Baskin II. The power of deep learning to ligand-based novel drug discovery. Expert Opin Drug Discov. 2020;15:755–64. https://doi.org/10.1080/17460441.2020.1745183
53. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA J Am Med Assoc. 2016;316:533–4. https://doi.org/10.1001/jama.2016.7653
54. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015;521:452–9. https://doi.org/10.1038/nature14541
55. Santosa F, Symes WW. Linear inversion of band-limited reflection seismograms. SIAM J Sci Stat Comput. 1986;7:1307–30. https://doi.org/10.1137/0907087
56. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Dent Tech. 1970;12:55–67. https://doi.org/10.1080/00401706.1970.10488634
57. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/bf00994018
58. Ignatov D, Ignatov A. Decision stream: cultivating deep decision trees. IEEE International Conference on Tools with Artificial Intelligence (ICTAI) 2017; 2017 November:905–12.
59. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324
60. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20:832–44. https://doi.org/10.1109/34.709601
61. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. Advances in Neural Information Processing Systems 2017; 2017 December:3857–67.
62. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front Environ Sci. 2016;3:80. https://doi.org/10.3389/fenvs.2015.00080
63. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. 2001;26.
64. Yang H, Sun L, Li W, Liu G, Tang Y. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front Chem. 2018;6:30. https://doi.org/10.3389/fchem.2018.00030
65. Hartwig A, Arand M, Epe B, Guth S, Jahnke G, Lampen A, et al. Mode of action-based risk assessment of genotoxic carcinogens. Arch Toxicol. 2020;94:1787–877. https://doi.org/10.1007/s00204-020-02733-2
66. Singh KP, Gupta S, Rai P. Predicting carcinogenicity of diverse chemicals using probabilistic neural network modeling approaches. Toxicol Appl Pharmacol. 2013;272:465–75. https://doi.org/10.1016/j.taap.2013.06.029
67. Fitzpatrick RB. CPDB: carcinogenic potency database. Med Ref Serv Q. 2008;27:303–11. https://doi.org/10.1080/02763860802198895
68. Zhu J, Chen J, Hu W, Zhang B. Big learning with Bayesian methods. Natl Sci Rev. 2017;4:627–51. https://doi.org/10.1093/nsr/nwx044
69. Guan D, Fan K, Spence I, Matthews S. Combining machine learning models of in vitro and in vivo bioassays improves rat carcinogenicity prediction. Regul Toxicol Pharmacol. 2018;94:8–15. https://doi.org/10.1016/j.yrtph.2018.01.008
70. GreenScreen® For Safer Chemicals | GreenScreen For Safer Chemicals® is an open, transparent, and publicly accessible method for chemical hazard assessment to help move our society quickly and effectively toward the use of greener and inherently safer chem; n.d.
71. Isfort RJ, Kerckaert GA, LeBoeuf RA. Comparison of the standard and reduced pH Syrian Hamster Embryo (SHE) cell in vitro transformation assays in predicting the carcinogenic potential of chemicals. Mutat Res Fundam Mol Mech Mutagen. 1996;356:11–63. https://doi.org/10.1016/0027-5107(95)00197-2
72. Hansen K, Mika S, Schroeter T, Sutter A, Ter Laak A, Thomas SH, et al. Benchmark data set for in silico prediction of Ames mutagenicity. J Chem Inf Model. 2009;49:2077–81. https://doi.org/10.1021/ci900161g
73. Benigni R, Bossa C, Richard AM, Yang C. A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity. Ann Ist Super Sanita. 2008;44:48–56.
74. Snyder RD. An update on the genotoxicity and carcinogenicity of marketed pharmaceuticals with reference to in silico predictivity. Environ Mol Mutagen. 2009;50:435–50. https://doi.org/10.1002/em.20485
75. Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Mach Learn. 1999;37:297–336. https://doi.org/10.1023/A:1007614523901
76. Zhu H, Martin TM, Ye L, Sedykh A, Young DM, Tropsha A. Quantitative structure–activity relationship modeling of rat acute toxicity by oral exposure. Chem Res Toxicol. 2009;22:1913–21. https://doi.org/10.1021/tx900189p
77. Xu Y, Pei J, Lai L. Deep learning based regression and multiclass models for acute Oral toxicity prediction with automatic chemical feature extraction. J Chem Inf Model. 2017;57:2672–85. https://doi.org/10.1021/acs.jcim.7b00244
78. AOT Prediction Server; n.d. Available from: http://www.pkumdl.cn:8080/DLAOT/DLAOThome.php. Accessed on 21 May 2021.
79. Li X, Kleinstreuer NC, Fourches D. Hierarchical quantitative structure–activity relationship modeling approach for integrating binary, multiclass, and regression models of acute oral systemic toxicity. Chem Res Toxicol. 2020;33:353–66. https://doi.org/10.1021/acs.chemrestox.9b00259

80. Alassane-Kpembi I, Gerez JR, Cossalter AM, Neves M, Laffitte J, Naylies C, et al. Intestinal toxicity of the type B trichothecene mycotoxin fusarenon-X: whole transcriptome profiling reveals new signaling pathways. Sci Rep. 2017;7:1–14. https://doi.org/10.1038/s41598-017-07155-2
81. Watson A, Opresko D, Young RA, Hauschild V, King J, Bakshi K. Organophosphate nerve agents. Handbook of toxicology of chemical warfare agents. 2nd ed. Oxford: Elsevier; 2015. p. 87–109. https://doi.org/10.1016/B978-0-12-800159-2.00009-9
82. Alberga D, Trisciuzzi D, Mansouri K, Mangiatordi GF, Nicolotti O. Prediction of acute oral systemic toxicity using a multifingerprint similarity approach. Toxicol Sci. 2019;167:484–95. https://doi.org/10.1093/toxsci/kfy255
83. Cao DS, Xu QS, Hu QN, Liang YZ. ChemoPy: freely available python package for computational biology and chemoinformatics. Bioinformatics. 2013;29:1092–4. https://doi.org/10.1093/bioinformatics/btt105
84. Zhang Y, Zhao J, Wang Y, Fan Y, Zhu L, Yang Y, et al. Prediction of hERG K+ channel blockage using deep neural networks. Chem Biol Drug Des. 2019;94:1973–85. https://doi.org/10.1111/cbdd.13600
85. Chavan S, Abdelaziz A, Wiklander JG, Nicholls IA. A k-nearest neighbor classification of hERG K+ channel blockers. J Comput Aided Mol Des. 2016;30:229–36. https://doi.org/10.1007/s10822-016-9898-z
86. Choi K-E, Balupuri A, Kang NS. The study on the hERG blocker prediction using chemical fingerprint analysis. Molecules. 2020;25:2615. https://doi.org/10.3390/molecules25112615
87. Ryu JY, Lee MY, Lee JH, Lee BH, Oh KS. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics. 2020;36:3049–55. https://doi.org/10.1093/bioinformatics/btaa075
88. Lee HM, Yu MS, Kazmi SR, Oh SY, Rhee KH, Bae MA, et al. Computational determination of hERG-related cardiotoxicity of drug candidates. BMC Bioinform. 2019;20:250. https://doi.org/10.1186/s12859-019-2814-5
89. Negami T, Araki M, Okuno Y, Terada T. Calculation of absolute binding free energies between the hERG channel and structurally diverse drugs. Sci Rep. 2019;9:1–12. https://doi.org/10.1038/s41598-019-53120-6
90. Beattie KA, Luscombe C, Williams G, Munoz-Muriedas J, Gavaghan DJ, Cui Y, et al. Evaluation of an in silico cardiac safety assay: using ion channel screening data to predict QT interval changes in the rabbit ventricular wedge. J Pharmacol Toxicol Methods. 2013;68:88–96. https://doi.org/10.1016/j.vascn.2013.04.004
91. Li Q, Jørgensen FS, Oprea T, Brunak S, Taboureau O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. Mol Pharm. 2008;5:117–27. https://doi.org/10.1021/mp700124e
92. Thai KM, Ecker GF. A binary QSAR model for classification of hERG potassium channel blockers. Bioorganic Med Chem. 2008;16:4107–19. https://doi.org/10.1016/j.bmc.2008.01.017
93. krictai / chemtrans — Bitbucket; n.d. Available from: https://bitbucket.org/krictai/chemtrans/src/master/. Accessed 14 May 2021.
94. kaistsystemsbiology / deepec — Bitbucket; n.d. Available from: https://bitbucket.org/kaistsystemsbiology/deepec/src/master/. Accessed 14 May 2021.
95. Wang Y, Huang L, Jiang S, Wang Y, Zou J, Fu H, et al. Capsule networks showed excellent performance in the classification of hERG blockers/nonblockers. Front Pharmacol. 2020;10:1. https://doi.org/10.3389/fphar.2019.01631
96. Saxena P, Zangerl-Plessl EM, Linder T, Windisch A, Hohaus A, Timin E, et al. New potential binding determinant for hERG channel inhibitors. Sci Rep. 2016;6:1–10. https://doi.org/10.1038/srep24182
97. Bian Y, Xie X-Q. Generative chemistry: drug discovery with deep learning generative models. J Mol Model. 2021;27:1–18. https://doi.org/10.1007/S00894-021-04674-8
98. Xue D, Gong Y, Yang Z, Chuai G, Qu S, Shen A, et al. Advances and challenges in deep generative models for de novo molecule generation. WIREs Comput Mol Sci. 2019;9:e1395. https://doi.org/10.1002/WCMS.1395
99. Prykhodko O, Johansson SV, Kotsias P-C, Arús-Pous J, Bjerrum EJ, Engkvist O, et al. A de novo molecular generation method using latent vector based generative adversarial network. J Cheminformatics. 2019;11:1–13. https://doi.org/10.1186/S13321-019-0397-9
100. Skalic M, Jiménez J, Sabbadin D, De Fabritiis G. Shape-based generative modeling for de novo drug design. J Chem Inf Model. 2019;59:1205–14. https://doi.org/10.1021/ACS.JCIM.8B00706
101. Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC); 2017. https://doi.org/10.26434/CHEMRXIV.5309668.V3
102. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2011;4:90–8. https://doi.org/10.1038/nchem.1243
103. Bai Q, Tan S, Xu T, Liu H, Huang J, Yao X. MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. Brief Bioinform. 2021;22:1–12. https://doi.org/10.1093/BIB/BBAA161
104. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24:232–40. https://doi.org/10.1093/bioinformatics/btn162
105. Hudson IL. Data integration using advances in machine learning in drug discovery and molecular biology. Methods in molecular biology. Volume 2190. Totowa, NJ: Humana Press; 2021. p. 167–84. https://doi.org/10.1007/978-1-0716-0826-5_7
106. Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform. 2019;93:103159. https://doi.org/10.1016/j.jbi.2019.103159
107. Jin X, Zhu L, Xue B, Zhu X, Yan D. Supramolecular nanoscale drug-delivery system with ordered structure. Natl Sci Rev. 2019;6:1128–37. https://doi.org/10.1093/nsr/nwz018
108. Chen R, Liu X, Jin S, Lin J, Liu J. Machine learning for drug–target interaction prediction. Molecules. 2018;23:1–15. https://doi.org/10.3390/molecules23092208

109. Rayhan F, Ahmed S, Mousavian Z, Farid DM, Shatabda S. FRnet-DTI: deep convolutional neural network for drug–target interaction prediction. Heliyon. 2020;6:e03444. https://doi.org/10.1016/j.heliyon.2020.e03444
110. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug–target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15:e1007129. https://doi.org/10.1371/journal.pcbi.1007129
111. Pliakos K, Vens C. Drug–target interaction prediction with tree-ensemble learning and output space reconstruction. BMC Bioinform. 2020;21:1V. https://doi.org/10.1186/s12859-020-3379-z
112. Li Y, Huang YA, You ZH, Li LP, Wang Z. Drug–target interaction prediction based on drug fingerprint information and protein sequence. Molecules. 2019;24:1–13. https://doi.org/10.3390/molecules24162999
113. Thafar MA, Thafar MA, Olayan RS, Olayan RS, Ashoor H, Ashoor H, et al. DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. J Cheminformatics. 2020;12:44. https://doi.org/10.1186/s13321-020-00447-2
114. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. Bioinformatics. 2018;34:1164–73. https://doi.org/10.1093/bioinformatics/btx731
115. Wang H, Wang J, Dong C, Lian Y, Liu D, Yan Z. A novel approach for drug–target interactions prediction based on multimodal deep autoencoder. Front Pharmacol. 2020;10:1592. https://doi.org/10.3389/fphar.2019.01592
116. Gerdes H, Casado P, Dokal A, Hijazi M, Akhtar N, Osuntola R, et al. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. Nat Commun. 2021;12:1–15. https://doi.org/10.1038/s41467-021-22170-8
117. Sonthalia S, Agrawal M, Sehgal VN, Gandhi V, Gupta KS. Drugs, discovery, and dermatology: Renbök, research and repurposing. Drug Discov Today. 2020;25:259–62. https://doi.org/10.1016/j.drudis.2019.10.011
118. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2018;18:41–58. https://doi.org/10.1038/nrd.2018.168
119. Shameer K, Glicksberg BS, Hodos R, Johnson KW, Badgeley MA, Readhead B, et al. Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. Brief Bioinform. 2018;19:656–78. https://doi.org/10.1093/bib/bbw136
120. Capecchi A, Awale M, Probst D, Reymond J. PubChem and ChEMBL beyond Lipinski. Mol Inform. 2019;38:1900016. https://doi.org/10.1002/minf.201900016
121. Bühlmann S, Reymond J-L. ChEMBL-likeness score and database GDBChEMBL. Front Chem. 2020;8:46. https://doi.org/10.3389/fchem.2020.00046
122. Brown AS, Patel CJ. A standard database for drug repositioning. Sci Data. 2017;4:1–7. https://doi.org/10.1038/sdata.2017.29
123. Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, et al. PubChem BioAssay: 2017 update. Nucleic Acids Res. 2017;45:D955–63. https://doi.org/10.1093/nar/gkw1118
124. Gaulton A, Hersey A, Nowotka ML, Patricia Bento A, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45:D945–54. https://doi.org/10.1093/nar/gkw1074
125. Liang L, Ma C, Du T, Zhao Y, Zhao X, Liu M, et al. Bioactivity-explorer: a web application for interactive visualization and exploration of bioactivity data. J Cheminformatics. 2019;11:47. https://doi.org/10.1186/s13321-019-0370-7
126. Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M, Masoudi-Nejad A. Drug databases and their contributions to drug repurposing. Genomics. 2020;112:1087–95. https://doi.org/10.1016/j.ygeno.2019.06.021
127. Zhao K, So HC. Using drug expression profiles and machine learning approach for drug repurposing. Methods in molecular biology. Volume 1903. Totowa, NJ: Humana Press; 2019. p. 219–37. https://doi.org/10.1007/978-1-4939-8955-3_13
128. Patrick MT, Raja K, Miller K, Sotzen J, Gudjonsson JE, Elder JT, et al. Drug repurposing prediction for immune-mediated cutaneous diseases using a word-embedding–based machine learning approach. J Invest Dermatol. 2019;139:683–91. https://doi.org/10.1016/j.jid.2018.09.018
129. Anderson E, Havener TM, Zorn KM, Foil DH, Lane TR, Capuzzi SJ, et al. Synergistic drug combinations and machine learning for drug repurposing in chordoma. Sci Rep. 2020;10:12982. https://doi.org/10.1038/s41598-020-70026-w
130. Liang S, Yu H. Revealing new therapeutic opportunities through drug target prediction: a class imbalance-tolerant machine learning approach. Bioinformatics. 2020;36:4490–7. https://doi.org/10.1093/bioinformatics/btaa495
131. Tanori M, Pannicelli A, Pasquali E, Casciati A, Antonelli F, Giardullo P, et al. Cancer risk from low dose radiation in Ptch1 +/− mice with inactive DNA repair systems: therapeutic implications for medulloblastoma. DNA Repair (Amst). 2019;74:70–9. https://doi.org/10.1016/j.dnarep.2018.12.003
132. Issa NT, Stathias V, Schürer S, Dakshanamurthy S. Machine and deep learning approaches for cancer drug repurposing. Semin Cancer Biol. 2021;68:132–42. https://doi.org/10.1016/j.semcancer.2019.12.011
133. Wang D, Liu W, Shen Z, Jiang L, Wang J, Li S, et al. Deep learning based drug metabolites prediction. Front Pharmacol. 2020;10:1586. https://doi.org/10.3389/fphar.2019.01586
134. Schneider G. Mind and machine in drug design. Nat Mach Intell. 2019;1:128–30. https://doi.org/10.1038/s42256-019-0030-7

How to cite this article: Staszak M, Staszak K, Wieszczycka K, Bajek A, Roszkowski K, Tylkowski B. Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. WIREs Comput Mol Sci. 2021;e1568. https://doi.org/10.1002/wcms.1568