Molecular Therapy
Nucleic Acids
Review

Artificial intelligence for drug discovery: Resources, methods, and applications

Wei Chen,[1,2] Xuesong Liu,[3] Sanyin Zhang,[1,2] and Shilin Chen[1,2]

[1]State Key Laboratory of Southwestern Chinese Medicine Resources, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China; [2]Institute of Herbgenomics, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China; [3]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

Conventional wet laboratory testing, validations, and synthetic procedures are costly and time-consuming for drug discovery. Advancements in artificial intelligence (AI) techniques have revolutionized their applications to drug discovery. Combined with accessible data resources, AI techniques are changing the landscape of drug discovery. In the past decades, a series of AI-based models have been developed for various steps of drug discovery. These models have been used as complements of conventional experiments and have accelerated the drug discovery process. In this review, we first introduced the widely used data resources in drug discovery, such as ChEMBL and DrugBank, followed by the molecular representation schemes that convert data into computer-readable formats. Meanwhile, we summarized the algorithms used to develop AI-based models for drug discovery. Subsequently, we discussed the applications of AI techniques in pharmaceutical analysis including predicting drug toxicity, drug bioactivity, and drug physicochemical property. Furthermore, we introduced the AI-based models for de novo drug design, drug-target structure prediction, drug-target interaction, and binding affinity prediction. Moreover, we also highlighted the advanced applications of AI in drug synergism/antagonism prediction and nanomedicine design. Finally, we discussed the challenges and future perspectives on the applications of AI to drug discovery.

INTRODUCTION

Drug discovery is a process through which new medications against diseases are discovered. It involves the use of a wide variety of technologies and expertise. In general, discovering and developing a drug takes US$2.8 billion and 15 years on average.[1] The low-efficacy and high-cost characteristics of conventional methods have become the hurdles of drug discovery. Therefore, developing new methods to deal with such a time-consuming and expensive task is necessary.[2]

The revolution in high-performance computer hardware and the availability of multi-omics data have enabled artificial intelligence (AI) techniques to transcend from theoretical studies to real applications in multiple disciplines. The successful application of AI techniques, particularly to biological data analysis, has attracted the attention of the pharmaceutical industry. Thus far, AI techniques have been implemented in drug discovery processes, such as drug-target prediction,[3] bioavailability prediction,[4] and *de novo* drug design.[5] Some major pharmaceutical companies, such as Bayer, Roche, and Pfizer, have also begun to collaborate with information technology (IT) companies to develop AI technique-based methods for drug design.[6] Recently, with the help of AI, the Insilico Medicine company discovered the drug treating idiopathic pulmonary fibrosis, which exhibits positive results in Phase I trials (https://clinicaltrials.gov/ct2/show/NCT05154240). Hence, drawing the conclusion that AI techniques have modernized the field of drug discovery and development is reasonable.

The basic schematics of applying AI techniques to drug discovery and evaluation are summarized in Figure 1. The major procedures include data collection and curation (Figure 1A), compound representation (Figure 1B), and AI methods and their applications in drug discovery (Figure 1C). To provide researchers with a catching-up view of the development in this field, we first introduced representative data resources, molecular representations and descriptors, and AI techniques in drug discovery. Then, we introduced the successful applications of AI to different stages of drug discovery. Finally, we discussed the challenges and future perspectives on applying AI to drug discovery.

RESOURCES AND METHODS FOR AI-BASED DRUG DISCOVERY

As indicated in Figure 1, data resources, data representation schemes, and AI methods are the three key components of applying AI to drug discovery and evaluation. Accordingly, they will be introduced briefly in this section.

Data resources

A high-quality dataset is the key to applying AI to drug discovery. Advances in high-throughput sequencing and IT have boosted the

https://doi.org/10.1016/j.omtn.2023.02.019.

Correspondence: Wei Chen, State Key Laboratory of Southwestern Chinese Medicine Resources, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China.
E-mail: greatchen@ncst.edu.cn
Correspondence: Shilin Chen, State Key Laboratory of Southwestern Chinese Medicine Resources, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China.
E-mail: slchen@icmm.ac.cn

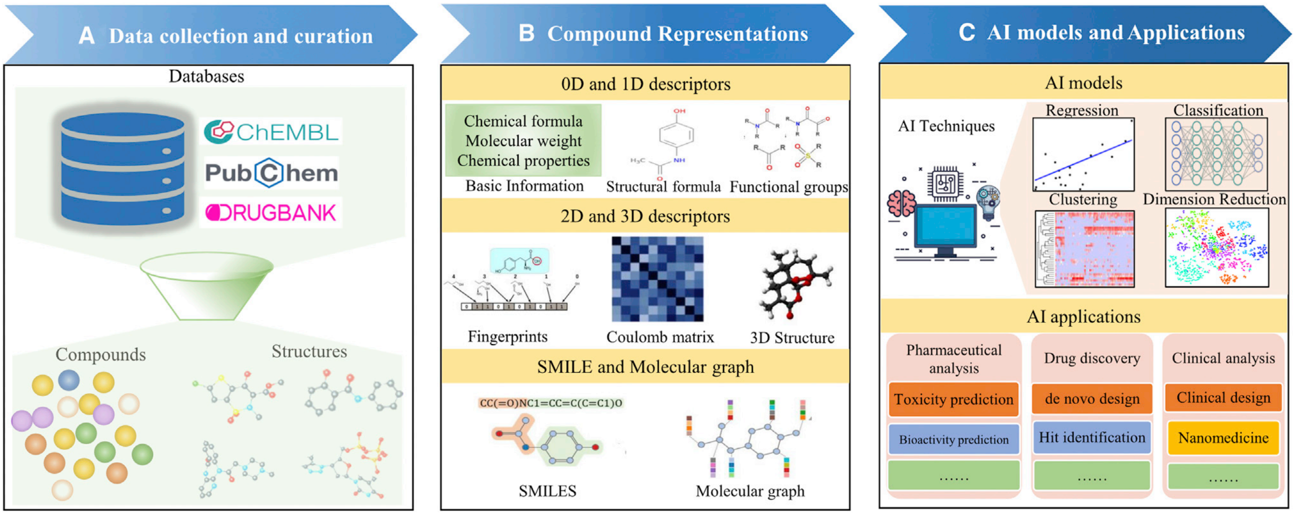

Figure 1. Framework of AI technique application to drug discovery and evaluation
Major procedures include (A) data collection and curation; (B) compound representations by using molecular descriptors; and (C) AI methods and their applications.

generation of a series of free and open-access databases for drug discovery. These databases enable drug discovery to transit into the big data era and accelerate the drug discovery process. Representative databases, along with their web links, brief descriptions, and references, are listed in Table 1. Their applications are not reviewed in the present work due to the limited space.

ChEMBL is a manually curated database that currently contains more than 2 million compounds that exhibit drug-like properties.[7] ChEMBL gathers information regarding the action mechanisms, molecular properties, absorption, distribution, metabolism, excretion, toxicity, therapeutic indications, and target interactions of the deposited compounds.

ChemDB is a freely accessible database that contains nearly 5 million commercially available small molecules and their physicochemical properties, such as molecular weight, solubility, and rotatable bonds.[8] In addition, a series of cheminformatics tools, such as Smi2Depict, MOLpro, AquaSol, and Reaction Predictor, are embedded into ChemDB, making this database user-friendly for drug discovery.

The Collection of Open Natural Products (COCONUT) is one of the best annotated databases of natural products.[9] It aggregates 407,270 elucidated and predicted natural products collected from a large number of chemical data sources. As a free database, COCONUT can be searched in multiple ways, such as molecule names, molecular structures, and structural properties. COCONUT also provides molecular properties and descriptors for each natural product. Moreover, all the data in COCONUT are available for download and can be queried programmatically via an application programming interface (API).

The Drug-Gene Interaction Database (DGIdb) provides information on drug-gene interactions and genes or gene products that can interact with drugs.[10] To date, DGIdb contains more than 40,000 genes and 10,000 drugs involved in over 100,000 drug-gene interactions. These data are mined from multiple diverse sources by performing expert curation and text mining. All the deposited genes in DGIdb are clustered into 43 categories. Users can either browse the genes in each category or enter a list of genes or drugs to retrieve drug-gene interactions in the search module. In addition, DGIdb can be accessed programmatically by API through the web-based interface.

DrugBank is a free-to-access reference drug database.[11] It currently contains 14,746 drugs, along with comprehensive information about drug-drug interactions, drug-target associations, drug classifications, and drug reactions. Users can search, browse, and extract text, images, and structural data in DrugBank by using the embedded tools. DrugBank has become the world's most widely used resource for drug screening, design, and metabolism prediction.

Drug Target Commons (DTC) is a freely accessible online resource that provides annotated and unannotated drug-target interaction (DTI) data.[12] For its recent release, DTC includes clinical trial information and disease-gene associations, facilitating the chemical biology and drug-repurposing applications of compounds. As an open resource, DTC not only supports database dump but also API to access its deposited data.

The Intelligent Network Pharmacology Platform Unique for Traditional Chinese Medicine (INPUT) is an online analytical platform that is uniquely for traditional Chinese medicine.[13] At present, INPUT contains 4,716 herbs, 29,812 herbal compounds, and 9,847 diseases collected from public databases and the literature. The herbs, compounds, and diseases are cross-linked through the herb-compound-gene-disease network in INPUT, which facilitates the

Table 1. Representative databases for drug discovery

Database	Website URL	Description	Reference
ChEMBL	https://www.ebi.ac.uk/chembl/	A manually curated database of bioactive molecules with drug-like properties. It gathers chemical, bioactivity, and genomic data to aid the translation of genomic information into effective new drugs.	Mendez et al.[7]
ChemDB	http://cdb.ics.uci.edu	A chemical database that contains nearly 5 million commercially available small molecules, along with their predicted or experimentally determined physicochemical properties.	Chen et al.[8]
COCONUT	https://coconut.naturalproducts.net/	A database that contains 407,270 unique natural products, along with information about their molecular properties and molecular descriptors.	Sorokina et al.[9]
DGIdb	http://www.dgidb.org	A database that provides information on DTI and druggable genomes from over 30 trusted sources.	Freshour et al.[10]
DrugBank	http://www.drugbank.ca	A database of drugs, their targets, 3D structures, and other useful information.	Wishart et al.[11]
DTC	http://drugtargetcommons.fimm.fi/	A crowd-sourcing platform that provides drug-target bioactivity data and classification of targets.	Tang et al.[12]
INPUT	http://cbcb.cdutcm.edu.cn/INPUT/	A network pharmacology platform for traditional Chinese medicine. It contains 29,812 compounds isolated from 4,716 Chinese herbs.	Li et al.[13]
PubChem	https://pubchem.ncbi.nlm.nih.gov/	An open chemistry database that provides information about molecules, such as chemical structures, identifiers, chemical and physical properties, and biological activities.	Kim et al.[14]
SIDER	http://sideeffects.embl.de	A database that provides information on marketed medicines and their recorded adverse reactions.	Campillos et al.[15]
STITCH	http://stitch.embl.de/	A database of known and predicted interactions between chemicals and proteins, including 9,643,763 proteins from 2,031 organisms.	Szklarczyk et al.[16]

discovery of herb-oriented drugs and the scientific interpretation of traditional Chinese medicine.

PubChem is a freely accessible chemical information resource that contains the biological, physical, chemical, and toxic information of chemical molecules.[14] All these data are collected from more than 850 sources. Users can search for chemicals in PubChem by inputting molecular formula, structure, and other identifiers as keywords. At present, PubChem has become one of the foremost data sources for computational drug discovery and design.

The Side Effect Resource (SIDER) is a database that focuses on drugs and their side effects.[15] The current release of SIDER includes 1,430 drugs, 5,880 side effects, and 140,064 drug-side effect pairs. These data can be browsed through either drugs or side effects. They have been used in many aspects, such as predicting drug indications, mining side effects, and identifying metabolic dysregulation.

The Search Tool for Interacting Chemicals (STITCH) is a database that contains known and predicted interactions between chemicals and proteins.[16] These interactions encompass 9,643,763 proteins from 2,031 organisms, which were collected from computational prediction, knowledge transfer between organisms, and other databases. Users can query STITCH in multiple ways, such as through the names of chemicals and proteins, chemical structures, and protein sequences. For large-scale analyses, the data in STITCH can be obtained either via bulk download or accessed programmatically with API.

Molecular descriptors and structure representations

With the explosive growth of natural products, another key point in AI-based drug discovery and analysis is the transfer of molecules into computer-readable format, while keeping their intrinsic physicochemical properties.[17] Various types of descriptors have been proposed to represent drugs; these descriptors can be classified into four categories in accordance with their dimensionality (Figure 2). To accelerate the drug discovery process, a series of open-source toolkits has been proposed for calculating molecular descriptors and structure representations, such as OpenBabel[18] and ChemmineR.[19]

The zero-dimensional (0D) descriptor is the simplest molecular representation; it is obtained in accordance with the chemical formula of drugs.[20] The 0D descriptor typically includes molecular weight, atom number, atom-type count, and other basic descriptors (e.g., number of heavy atoms). The 0D descriptor is extremely simple, and it can only extract shallow information.

The one-dimensional (1D) descriptor encodes drugs in accordance with their substructures, such as the number of rings, functional groups, substituent atoms, and atom-centered fragments.[20] The elements of the 1D descriptor are typically binary (e.g., 1/0 indicates the presence/absence of a substituent atom) or the occurrence frequencies of some substructures. Apart from the property-based 1D descriptor, the simplified molecular-input line-entry system (SMILES)[21] is another type of 1D descriptor. SMILES represents drugs with a string of characters. SMILES depends on atom order,

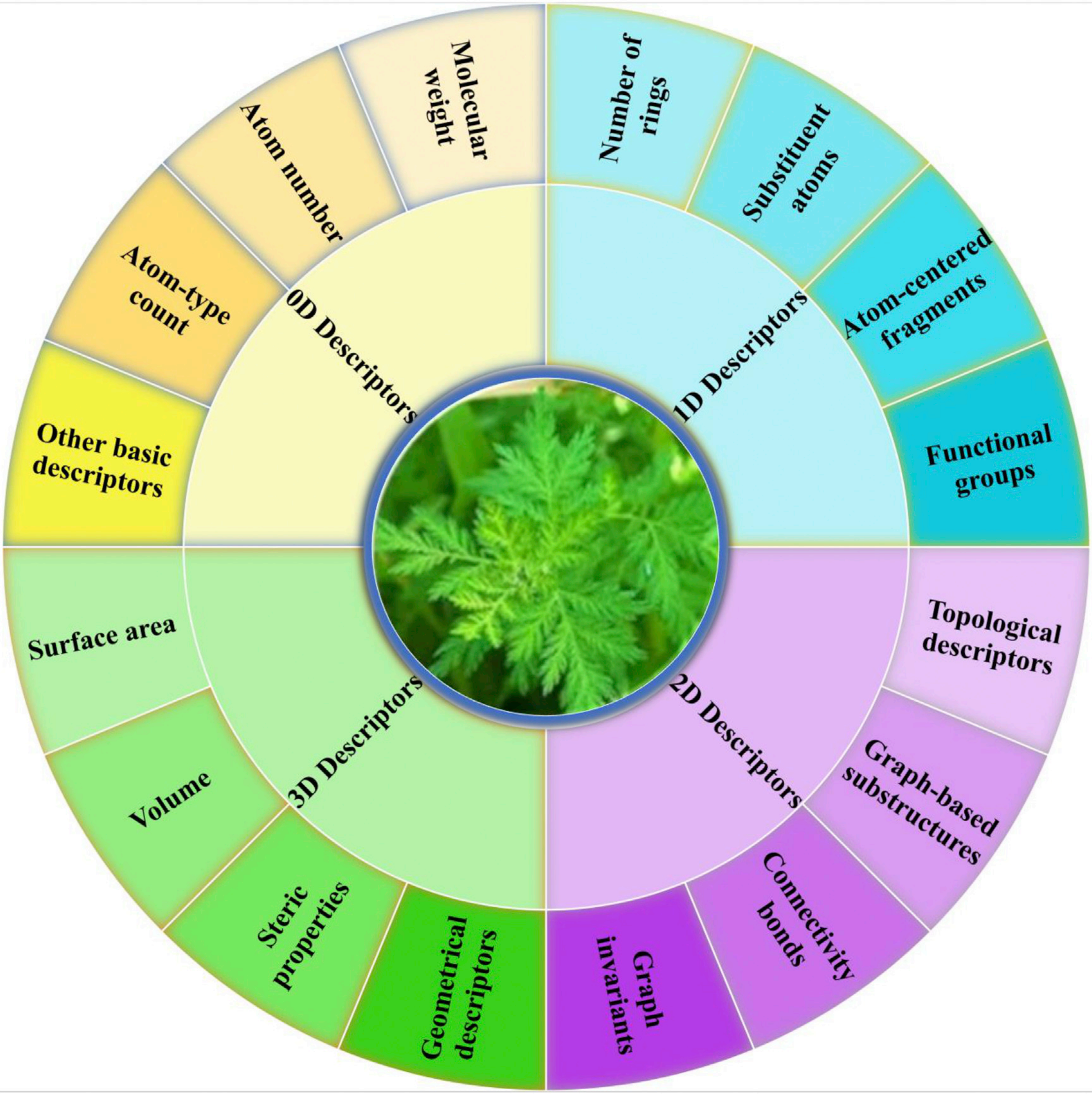

Figure 2. Summary of molecular and structural representation schemes

and thus, a drug will have several SMILES representations, and the normalization algorithm should be performed to obtain canonical SMILES.

The two-dimensional (2D) descriptor provides additional information to the 1D descriptor by considering adjacency, connectivity, and other types of topological features of the atoms. Therefore, 2D descriptors are typically derived by representing a drug as a graph, wherein the nodes indicate atoms and edges indicate bonds. Property-based 2D descriptors frequently include graph invariants, connectivity bonds, graph-based substructures, and topological descriptors. To extract more information, the molecular fingerprint (FP) was proposed for encoding molecules in binary form.[22] FP indicates the presence/absence of particular substructures through a string with a given length and marked by 1/0. The commonly used 2D FPs are the molecular access system fingerprints,[23] daylight-like fingerprint,[18] and extended-connectivity fingerprints.[24]

The three-dimensional (3D) descriptor depicts a molecule in 3D space,[25] and each atom of a molecule is spatially characterized by the x, y, and z coordinates. The 3D descriptor includes spatial and geometrical configuration information; it has high information content. Thus, information about surface area, volume, and steric properties can be obtained by using 3D descriptors. Non-property-based 3D descriptors, such as geometrical fingerprint[26] and pharmacophore fingerprint,[27] are also available. They can represent complex physicochemical properties of drugs and are widely used in drug discovery and virtual screening.

Table 2. Widely used AI techniques in drug discovery

Category	Task	Method	Representative application	Reference
Supervised learning	Regression analysis	MLR	DTI	Talevi et al.[29]
		DT	Adverse drug reactions	Hammann et al.[33]
		LR	Drug-drug interaction	Schober and Vetter[34]
	Classification	SVM	Compound classification	Maltarollo et al.[35]
		CNN	Bioactivity prediction	El-Attar et al.[36]
		RNN	*De novo* drug design	Gupta et al.[37]
		GAN	Molecule discovery	Blanchard et al.[38]
Unsupervised learning	Clustering	*k*-means	Drug candidate selection	Shen et al.[39]
		Hierarchical	Molecular scaffold analysis	Manelfi et al.[40]
	Dimension reduction	PCA	QSAR	Yoo and Shahlaei[41]
		t-SNE	Chemical space mapping	Karlov et al.[42]

CNN, convolution neural network; DT, drug target; GAN, generative adversarial network; LR, logistic regression; MLR, multiple linear regression; PCA, principal-component analysis; RNN, recurrent neural network; SVM, support vector machine; t-SNE, T-distributed stochastic neighbor embedding.

The schematic diagrams illustrating the representations of compounds by using 0D-3D descriptors are shown in Figure 1B. In addition to these encoding schemes, graph-based methods have also been proposed recently to encode molecules. Examples of the graph-based schemes include the spectral and spatial graph convolutional network. For more details about graph-based molecular representation methods, readers can refer to a recent review.[28]

Commonly used AI techniques

To be a fit-for-purpose approach, the selection and application of AI techniques are problem-oriented. Two common types of AI techniques, namely, supervised and unsupervised learning, are used in the field of drug discovery.[29] A supervised learning technique uses input-labeled data to train models that are capable of classifying or predicting outcomes of new data. By contrast, an unsupervised learning technique deals with unlabeled data and aims to develop models that are capable of identifying recurring patterns and clustering of the input data in a manner without prior knowledge.[30] Supervised learning techniques can be further classified into classification and regression algorithms, and unsupervised learning techniques include clustering and dimensionality reduction algorithms. To facilitate users in applying these AI techniques, a series of open-source packages and frameworks, such as Scikit-learn,[31] PyTorch,[32] and Keras (https://github.com/fchollet/keras), have been developed for practicing the aforementioned algorithms. Widely used AI techniques in drug discovery are listed in Table 2 and briefly discussed below.

Regression analysis technique

Multiple linear regression (MLR) is a modeling technique that aims to estimate the relationship between independent variables and the dependent variable by fitting a linear equation into observed data.[29] The ordinary least squares method is used to find the best-fit line by reducing the sum of squared errors, which are the differences between the observed value and the fitted value given by the model.

A decision tree (DT) is a nonlinear supervised learning technique that can be used in classification and regression tasks.[33] The primary components of a DT model are nodes (including root nodes, internal nodes, and leaf nodes) and branches. The algorithm starts at the root node and selects a branch according to the decision rule of the root node. Subsequently, the algorithm reaches the internal nodes and further makes decisions on the basis of this node. Finally, the algorithm will reach leaf nodes that represent possible outcomes within the dataset.

Logistic regression (LR) is a supervised learning technique that can be used to estimate the probability of occurrence of an event on the basis of log odds ratio.[34] LR can be classified into three categories, namely, binary, nominal, and ordinal LR, in accordance with the categories of response variables.

Classification technique

Support vector machine (SVM) is a classical supervised learning technique that is widely used in drug discovery.[35] The basic idea of SVM is to cast data into higher-dimensional feature space by using kernel functions and find the optimal separating hyperplane that maximizes the margin of training data.

Convolution neural network (CNN) is a deep learning technique with feedforward neural network architecture.[36] The CNN model includes three types of layers: the convolutional, pooling, and fully connected layers. The convolutional layer aims to learn feature representations of the input. The pooling layer is used to reduce the number of trainable parameters. The fully connected layer aims to produce classification scores and perform reasoning. Compared with conventional machine learning methods, the advantages of CNN include automatically extracting non-handcrafter features from raw input.

Recurrent neural network (RNN) is a feedforward artificial neural network (ANN) that specializes in dealing with sequential data.[37]

RNN consists of numerous successive recurrent layers, and its information cycles through a loop. These features make RNN distinct from the traditional neural network. Hence, RNN has the ability to capture contextual content from input data. RNN has also been used in drug design and discovery given its great promise in handling sequential data.[43]

Generative adversarial network (GAN) is a deep learning framework with two components: the generator and the discriminator.[38] The former is used to generate new data with the same characteristics as the training data. The latter is used to distinguish actual samples from the generated fake ones. Compared with conventional machine learning methods and other deep learning techniques, GAN is good at solving problems with a small sample size.

Clustering technique

k-means clustering is one of the most important and popular clustering algorithms.[39] It aims to group similar data into clusters, such that samples in the same cluster are more similar to each other than to those in other clusters. This algorithm iteratively identifies a certain number of centroids (i.e., the arithmetic mean of all data points assigned to a particular cluster) within a dataset and allocates every datum to the nearest cluster. These procedures are repeated until cluster assignments stop changing.

Hierarchical clustering is another type of clustering algorithm that is used to group data into clusters on the basis of similarity measures.[40] Distinct from *k*-means clustering, hierarchical clustering initially regards each datum as an individual cluster and then identifies the two closest clusters and merges them together. These procedures are iterated until all the clusters are merged together. The final result is presented in a dendrogram.

Dimension reduction

Principal-component analysis (PCA) is a linear dimensionality reduction technique that can transform a large dataset into a smaller one while maintaining most of the original information.[41] The basic idea of PCA is to find principal components that explain a large portion of the variation in a dataset. The procedures for conducting PCA include standardizing data, computing the covariance matrix, computing the eigenvalues and eigenvectors, identifying the principal components, and remodeling the data.

T-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction technique that is capable of visualizing high-dimensional data in 2D or 3D space.[42] The t-SNE algorithm first converts similarities between data points into joint probabilities. Then, it minimizes the Kullback-Leibler divergence between the joint probabilities of high-dimensional data and low-dimensional embedding.

Application of AI to pharmaceutical analysis

Pharmaceutical analysis involves the processes of identification, determination, quantification, and purification of pharmaceutical raw materials; it is an essential part of drug discovery. Qualitative and quantitative analyses are the two major types of experimental methods in pharmaceutical analysis. Although these techniques exhibit high accuracy, their cost for screening novel drug candidates from a huge amount of natural products is still expensive. Compared with experimental techniques, the costs required by computational methods are negligible. Hence, AI techniques have been used in pharmaceutical analysis to complement experimental techniques. The representative applications of AI techniques in pharmaceutical analysis are summarized in Figure 3.

Drug toxicity prediction

Toxicity is a measure of the unwanted or adverse effects of chemicals.[44] Toxicity evaluation is one of the fundamental steps in drug discovery, and it aims to identify substances that have harmful effects on humans.[45] However, the *in vivo* test requires animal tests and thus increases the costs of drug discovery. Computational methods exhibit the advantages of being able to predict a chemical's toxicity with low cost and high efficiency.[46] Accordingly, a series of AI technique-based methods have been developed to predict the toxicity of chemicals.[47,48] To assess the performance of different computational methods for predicting the toxicity of chemicals, the scientific community proposed the "Toxicology in the 21st Century (Tox21)" challenge.[46]

DeepTox is an ensemble model for predicting the toxicity of chemicals, and its fundamental framework is based on a three-layer deep neural network (DNN).[49] After performing data cleaning and quality control, the remaining chemicals are encoded by using the aforementioned 0D to 3D molecular descriptors, which are used as input of DNN. The DeepTox pipeline is obtained by tuning and optimizing a set of hyperparameters, such as number of hidden units, learning rate, and dropout rate. Comparative results based on the Tox21 dataset demonstrate that DeepTox outperforms its counterparts in toxicity prediction.[49]

Drug bioactivity prediction

In reality, a large number of drugs derived from natural products are ineffective due to the lack of bioactivity. Hence, drug bioactivity assessment has become an active area in drug discovery. Although *in vitro* and *in vivo* experiments can mimic the functions of molecules in the human body, they are still time-consuming and expensive. Given their cost-effectiveness and time economy, AI techniques have been effectively applied to predicting drug bioactivities, such as anticancer, antiviral, and antibacterial activities.[50–52]

For example, Stokes et al. proposed a directed message passing neural network that is capable of predicting antibacterial activity.[53] For each molecule, they first constructed a molecular graph in accordance with its SMILES and then obtained the feature vector based on atomic features (e.g., number of bonds for each atom and atomic number) and bond features (e.g., bond type and stereochemistry).[53] By applying the message passing operation multiple times, the optimized feature vector was fed into the feedforward neural network that outputted the antibacterial probability of a molecule.[53] This model

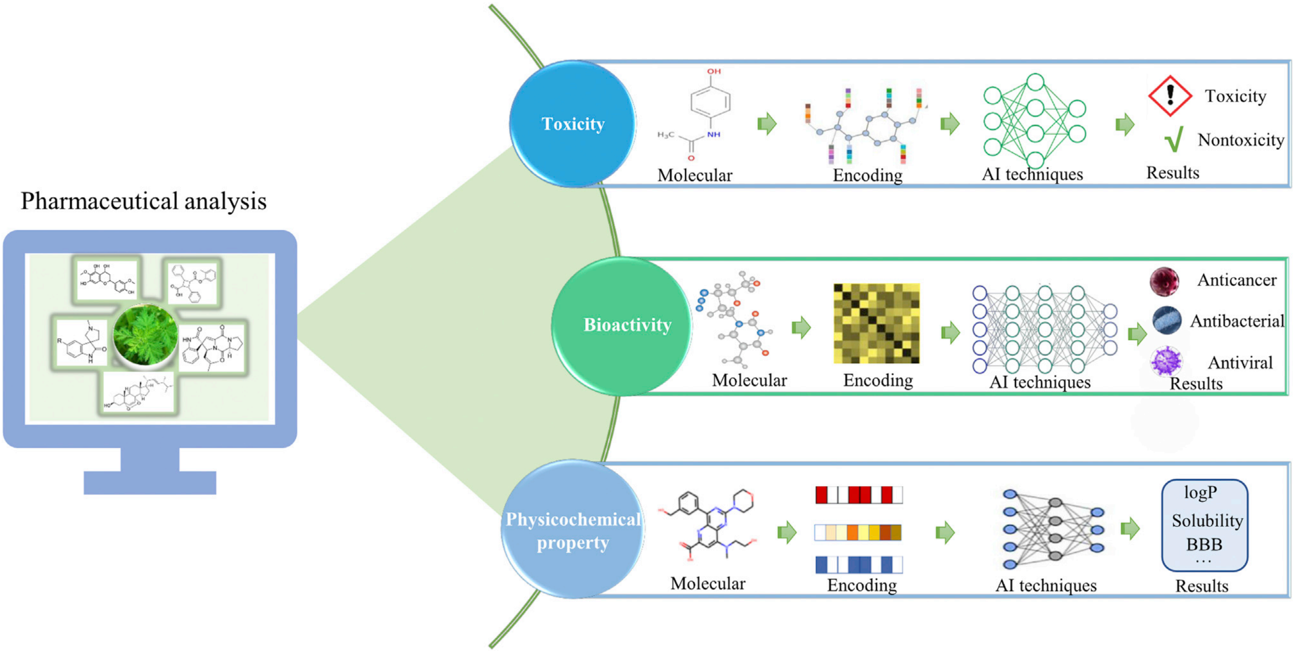

Figure 3. Application of AI techniques to pharmaceutical analysis

is available at http://chemprop.csail.mit.edu/, and it can facilitate the discovery of antibacterial molecules.

Drug physicochemical property prediction

Physicochemical properties are intrinsic characteristics of drugs. Knowledge about physicochemical properties is required for understanding and modeling the action of drugs. Among the numerous types of physicochemical properties, solubility is important because it affects the pharmacokinetic properties and formulations of drugs.[54,55] However, laborious and costly experimental techniques have precluded rapid solubility prediction; hence, considerable effort has been devoted to develop AI-based solubility prediction models.

Panapitiya et al. assessed different deep learning methods (i.e., fully connected neural networks, RNNs, graph neural networks, and SchNet) and molecular representation approaches (i.e., molecular descriptors, SMILES, molecular graphs, and 3D atomic coordinates) for solubility prediction.[54] Based on the same test dataset, the authors found that the fully connected neural network achieved the best performance for solubility prediction by leveraging molecular descriptors. In addition, the authors analyzed the importance of different features for prediction and found that 2D molecular descriptors made the greatest contributions. To facilitate further research on solubility prediction, an open-source code was provided at https://github.com/pnnl/solubility-prediction-paper.

AI in natural product-inspired drug discovery

Drug discovery is a process of identifying active compounds with therapeutic effects on the intended diseases. Although a high-throughput screening technique can scan thousands of different compounds one at a time, it is still time-consuming and costly.[56] To address these challenges, AI techniques have been applied to nearly all aspects of drug discovery. The applications of AI to natural product-inspired drug discovery, such as *de novo* drug design, target structure prediction, DTI prediction, and drug-target binding affinity prediction, are illustrated in Figure 4.

De novo drug design

De novo drug design refers to the process of generating novel drug-like compounds without a starting template. Although conventional structure-based and ligand-based drug design methods have enhanced the discovery of small-molecule drug candidates, they respectively rely on knowledge about the active site of a biological target or the pharmacophores of a known active binder,[57] hindering their applications to modern drug discovery. The boom of AI techniques has offered new opportunities to *de novo* drug design and accelerated the drug discovery process.

In recent years, various deep learning-based models have been proposed for *de novo* drug design, such as the reinforcement learning-based model ReLeaSE,[58] the encoder-decoder-based model ChemVAE,[59] the GAN-based model GraphINVENT,[60] and the RNN-based model MolRNN.[61] Another key point of *de novo* drug design is molecular representation. SMILES, fingerprint, molecular graph, and 3D geometry have been used as input of deep learning algorithms. The fundamental framework of deep learning-based *de novo* drug design methods is shown in the left upper corner of

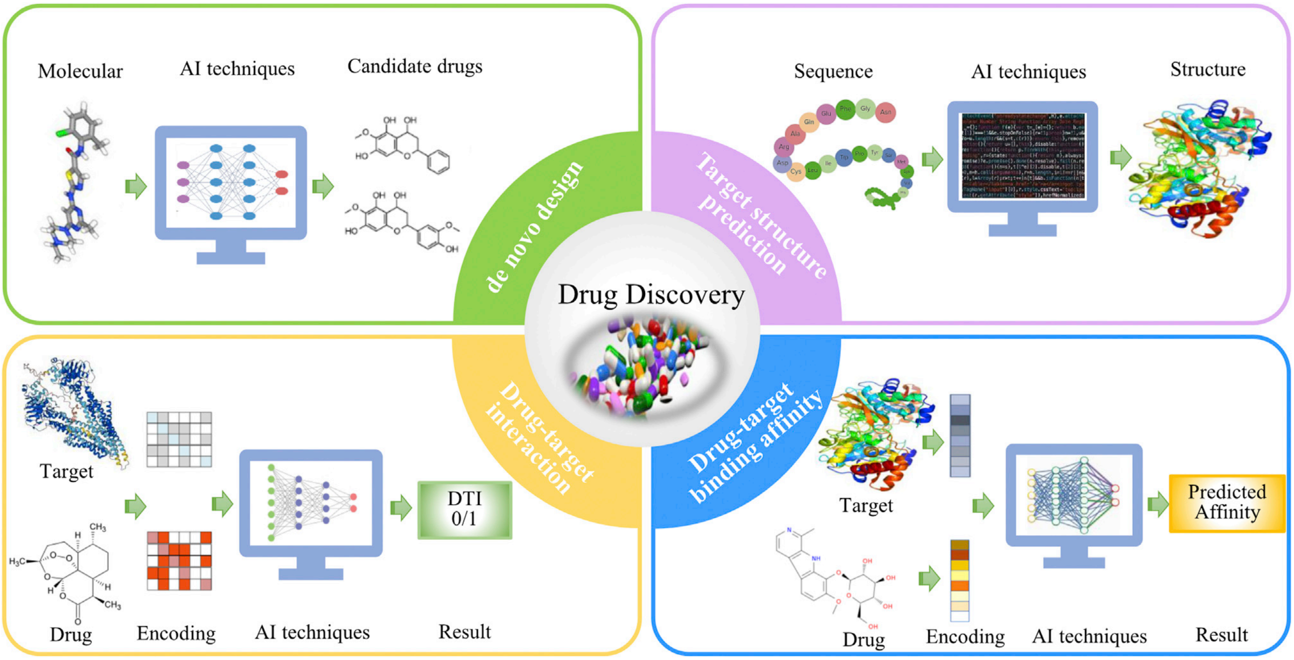

Figure 4. AI techniques for natural product-inspired drug discovery

Figure 4. Detailed information about deep learning-based *de novo* drug design models is provided recent reviews.[57,62]

Target structure prediction

Most drug targets are proteins that play important roles in enzymatic activities, cell signaling, and cell-cell transduction. The functions of proteins are determined by their structures. Although conventional experimental techniques, such as X-ray crystallography, cryogenic electron microscopy, and nuclear magnetic resonance spectroscopy, have been proposed to determine protein structures, they are still time-consuming and costly.[63] As reported, experimental techniques have only deciphered the structures of 100,000 unique proteins, which account for only a small part of known proteins.[64] Therefore, developing novel methods to fill the gap between the number of protein sequences and known protein structures is an urgent need.[65]

With the rapid growth of computational power and the breakthroughs of AI techniques, many computational approaches have been proposed for protein structure prediction. The basic schematics of computational protein structure prediction models are presented in the right upper corner of Figure 4. The neural network-based AlphaFold method developed by DeepMind is the best-performing method, and it is able to predict the 3D structures of proteins from their amino acid sequences and achieve accuracies competitive with experiments.[64] The descriptions of the algorithm and architecture of AlphaFold are provided in Senior et al.[66] The source code of AlphaFold is available at https://github.com/deepmind/alphafold.

DTI prediction

DTI prediction refers to the interaction between chemical compounds and protein targets in living organisms.[67] DTI prediction is an essential process for drug discovery. Hence, experimental methods have been used to determine DTI, such as co-immunoprecipitation,[68] phage display technology,[69] and yeast two-hybrid.[70] However, these wet laboratory techniques are time-consuming when they are used to predict DTI. Recently, the ever-increasing biological data have paved the way for the *in silico* prediction of DTI. Therefore, computational methods are being increasingly used in DTI prediction. These methods, which were summarized in a recent review,[71] can be classified into the following categories: ligand-based methods, docking simulations, gene ontology-based methods, text mining-based methods, and network-based methods.

Compared with other types of methods, deep learning-based methods frequently exhibit better performance in DTI prediction.[72] The common workflow of the deep learning-based DTI prediction method is illustrated at the left bottom corner of Figure 4. First, compounds and proteins are encoded by using their corresponding features. Then, the feature embedding of the compounds and proteins is used as the input of deep learning methods. In accordance with this strategy, models based on deep belief neural network,[73] CNN,[72] and multiple layer perceptron[74] have been proposed for drug-protein interaction prediction, considerably facilitating drug discovery.

In real life, many diseases lack well-defined targets. Hence, finding drugs for these diseases is impossible by using the aforementioned methods. Zhu et al. recently proposed a deep learning-based efficacy prediction system (DLEPS) that can identify drug candidates in accordance with the changes in gene expression profiles rather than specific targets.[75] First, compounds were encoded using SMILES and used as input of CNN to fit gene expression changes. Subsequently, the potential efficacy of compounds against diseases was evaluated on the basis of gene signatures specific to certain diseases and sorted using a method similar to gene set enrichment analysis. DLEPS provides novel insights into identifying new drugs for complex diseases.

Drug-target binding affinity prediction

In most cases, DTI prediction is regarded as a binary classification problem, but binding affinity between a drug and its target is disregarded.[67] Binding affinity reflects the strength of drug-target pair interactions, and it is considerably informative for drug discovery. Although binding affinity can be experimentally determined by measuring dissociation and inhibition constants, the time cost and financial expenses of these procedures are extremely high. Therefore, developing computational methods for predicting binding affinity is necessary.

In 2018, Öztürk et al. proposed the first deep learning model, called DeepDTA, for predicting binding affinity between drugs and their targets.[76] In DeepDTA, the drug and the target were encoded using SMILES and amino acid letters, respectively, which were then used as input for CNN. The basic framework of DeepDTA is shown at the right bottom corner of Figure 4. The comparative results demonstrated that DeepDTA suppressed KronRLS[77] and SimBoost[78] for drug-target binding affinity prediction. Inspired by DeepDTA, a series of deep learning-based models has been sequentially proposed, such as WideDTA[76] and DeepAffinity,[79] which have become useful tools in drug discovery.

Advanced applications of AI in drug design

AI in drug synergism/antagonism prediction

Synergism and antagonism are the two categories of drug combination effects. The former can overcome primary and secondary drug resistance, and it is effective for the treatment of cancers,[80] AIDS,[81] and bacterial infections,[82] whereas the latter reduces the effectiveness of drugs. With the ever-increasing number of drugs, their possible combinations are astronomical. Thus, experimentally investigating drug combination effect is costly and time-consuming. The advancements of AI techniques have made them applicable to exploring possible drug combinations at lower cost and with more efficiency.

In 2015, Li et al. proposed a Bayesian network model for exploring and analyzing drug combinations.[83] In the same year, Wildenhain et al. developed a random forest-based model for predicting compound synergism from chemical-genetic interactions.[84] Recently, Preuer et al. proposed DeepSynergy,[85] a deep learning-based model for predicting the synergism of anticancer drugs. The inputs of DeepSynergy included the chemical information of drugs and the genomic information of diseases, which were then propagated through the network to the output unit. The comparative results from a publicly available synergy dataset demonstrated that DeepSynergy outperformed its counterparts in predicting drug synergism. The web server and source code of DeepSynergy are provided at www.bioinf.jku.at/software/DeepSynergy and https://github.com/KristinaPreuer/DeepSynergy, respectively.

AI in nanomedicine design

Nanotechnology has been applied to design nanomedicines by using nanometric-scale materials in the clinical setting.[86] Nanomedicines are developed by materials at the nanometric scale, and, thus, they can penetrate the barriers to interact with targets in the body. At present, some nanomedicines have already been approved by the U.S. Food and Drug Administration, and they have exhibited better performance in the treatment of cancers[87] and HIV-1 infection.[88] However, the lack of quantitative and qualitative understanding of nanomaterial properties and biological responses precluded the wide application of nanomedicines.

A combination of nanotechnology and AI provides novel solutions to deal with this dilemma. For example, Li et al. proposed an ANN for the task of nanomedicine composition optimization.[89] Muñiz Castro et al. developed a 3D printing nanomaterial formulation pipeline that can predict the extrusion temperature, filament mechanical characteristics, and dissolution time of nanomaterials.[90] In addition, the effectiveness of a nanomedicine is affected by its cellular uptake. Hence, a cellular uptake prediction model will considerably help researchers in predicting nanomedicine effectiveness. On the basis of an ANN, Alafeef et al. developed a platform for predicting nanoparticle cellular internalization in different cell types.[91] Other applications of AI to nanomedicine design and their principles were summarized in a recent comprehensive review.[80]

AI in oligonucleotide design

Besides the drugs derived from natural products, oligonucleotide therapeutics composed of short strands of DNA or RNA have become a novel class of drugs.[92] Antisense oligonucleotides (ASO), small interfering RNA (siRNA), and CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated protein) are the main oligonucleotide therapeutics systems that enable the precise treatment of diverse diseases. Since experimental designing these oligonucleotides will cost enormous resources, the AI approaches have also been used to help researchers to identify and design the oligonucleotide-based drugs. For example, Chiba et al. proposed a machine learning-based model, eSkip-Finder, to identify effective exon skipping ASOs.[93] Dar et al. developed SMEpred to predict the efficacy of siRNAs.[94]

CONCLUDING REMARKS AND PROSPECTS

Over the past few years, we have witnessed the wide applications of AI techniques to various steps of drug discovery and development. The boom of AI techniques has made substantial contributions to the acceleration of drug discovery. The application of Chat Generative Pre-Trained Transformer (ChatGPT) is also a promising topic in drug discovery and development. Since it can provide methods to identify potential targets, design new drugs, and optimize the pharmacodynamics of drug candidates, ChatGPT has the potential to speed up drug development process. However, AI techniques are not versatile tools for drug discovery due to the following challenges.

The first key point is the availability of high-quality data that can be used to train AI technique-based models. Although the amount of biological and chemical data is increasing, the issue of poor data quality hinders the full use of these data. To solve this issue, data curation can be performed to organize and manage raw data. For this objective, academic institutions and pharmaceutical companies should cooperate to develop data standards and frameworks that will be helpful in data collection and clearance. Data quantity is also important for the applications of AI techniques. In real cases, the number of positive samples is smaller than that of negative ones. The sample imbalance problem will directly affect the performance of the models. Thus, oversampling and undersampling methods are suggested to be used to balance samples in the datasets.

Another typical issue of AI technique-based models for drug discovery is the lack of interpretability. A model's interpretability is the degree to which humans can understand the processes it uses to arrive at its outcomes. In most cases, the proposed models fall short in interpreting their biological and pharmaceutical meanings. Hence, trusting the predictive results obtained by AI techniques is difficult for experimental scientists.[95] In addition, the lack of interpretability also makes models unable to troubleshoot these approaches when their performance is poor on the test data. To deal with this issue, post hoc explanation techniques are suggested to be used when building models.[96] Popular techniques for post hoc interpretations include text explanation, visualization explanation, and attention mechanism explanation. Text explanation techniques can provide qualitative interpretations by presenting human-understandable verbal words. Visualization explanation techniques, such as t-SNE, can visualize the learned latent high-dimensional features in 2D space.[96] Attention mechanism explanation techniques can automatically learn and calculate the contribution of input to output, making the model interpretable.[97]

The availability and accessibility of the proposed models are also challenges in drug discovery. Although many AI technique-based models have been developed, neither related freely accessible web servers nor source codes are provided for most of these models. Even though some smart tools have been designed, they are only commercially available. These issues preclude their applications to drug discovery and development. Hence, developing open-source tools or packages, which will become invaluable sources in the near future, is necessary.

Although there exist the above-mentioned challenges, AI techniques have been incorporated into drug discovery and development industry. It is believable that AI techniques will bring revolutionary changes for this field.

ACKNOWLEDGMENTS

This study was supported by Natural Science Foundation of Sichuan (No. 2023NSFSC0683) and Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (No: ZYYCXTD-D-202209).

AUTHOR CONTRIBUTIONS

S.C. and W.C. conceived the work, X.L and S.Z. participated in reviewing and revising the manuscript. S.C. and W.C. wrote and revised the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Fleming, N. (2018). How artificial intelligence is changing drug discovery. Nature *557*, S55–S57.
2. Chen, S., Li, Z., Zhang, S., Zhou, Y., Xiao, X., Cui, P., Xu, B., Zhao, Q., Kong, S., and Dai, Y. (2022). Emerging biotechnology applications in natural product and synthetic pharmaceutical analyses. Acta Pharm. Sin. B *12*, 4075–4097.
3. You, Y., Lai, X., Pan, Y., Zheng, H., Vera, J., Liu, S., Deng, S., and Zhang, L. (2022). Artificial intelligence in cancer target identification and drug discovery. Signal Transduct. Targeted Ther. *7*, 156.
4. Wei, M., Zhang, X., Pan, X., Wang, B., Ji, C., Qi, Y., and Zhang, J.Z.H. (2022). HobPre: accurate prediction of human oral bioavailability for small molecules. J. Cheminf. *14*, 1.
5. Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., Papadopoulos, K., and Patronov, A. (2020). Reinvent 2.0: an AI tool for de novo drug design. J. Chem. Inf. Model. *60*, 5918–5922.
6. Mak, K.K., and Pichika, M.R. (2019). Artificial intelligence in drug development: present status and future prospects. Drug Discov. Today *24*, 773–780.
7. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., et al. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. *47*, D930–D940.
8. Chen, J., Swamidass, S.J., Dou, Y., Bruand, J., and Baldi, P. (2005). ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics *21*, 4133–4139.
9. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M.A., and Steinbeck, C. (2021). COCONUT online: collection of open natural products database. J. Cheminf. *13*, 2.
10. Freshour, S.L., Kiwala, S., Cotto, K.C., Coffman, A.C., McMichael, J.F., Song, J.J., Griffith, M., Griffith, O.L., and Wagner, A.H. (2021). Integration of the drug-gene interaction database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res. *49*, D1144–D1151.
11. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.
12. Tang, J., Tanoli, Z.U.R., Ravikumar, B., Alam, Z., Rebane, A., Vähä-Koskela, M., Peddinti, G., van Adrichem, A.J., Wakkinen, J., Jaiswal, A., et al. (2018). Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. Cell Chem. Biol. *25*, 224–229.e2.
13. Li, X., Tang, Q., Meng, F., Du, P., and Chen, W. (2022). INPUT: an intelligent network pharmacology platform unique for traditional Chinese medicine. Comput. Struct. Biotechnol. J. *20*, 1345–1351.
14. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. *49*, D1388–D1395.
15. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., and Bork, P. (2008). Drug target identification using side-effect similarity. Science *321*, 263–266.
16. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. *49*, D605–D612.
17. David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. J. Cheminf. *12*, 56.

18. O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: an open chemical toolbox. J. Cheminf. *3*, 33.

19. Cao, Y., Charisi, A., Cheng, L.C., Jiang, T., and Girke, T. (2008). ChemmineR: a compound mining framework for R. Bioinformatics *24*, 1733–1734.

20. Grisoni, F., Ballabio, D., Todeschini, R., and Consonni, V. (2018). Molecular descriptors for structure-activity applications: a hands-on approach. Methods Mol. Biol. *1800*, 3–53.

21. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model. *28*, 31–36.

22. Capecchi, A., Probst, D., and Reymond, J.L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J. Cheminf. *12*, 43.

23. Seo, M., Shin, H.K., Myung, Y., Hwang, S., and No, K.T. (2020). Development of natural compound molecular fingerprint (NC-MFP) with the dictionary of natural products (DNP) for natural product-based drug development. J. Cheminf. *12*, 6.

24. Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754.

25. Matter, H., and Pötter, T. (1999). Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. J. Chem. Inf. Comput. Sci. *39*, 1211–1225.

26. Yin, S., Proctor, E.A., Lugovskoy, A.A., and Dokholyan, N.V. (2009). Fast screening of protein surfaces using geometric invariant fingerprints. Proc. Natl. Acad. Sci. USA *106*, 16622–16626.

27. Wood, D.J., de Vlieg, J., Wagener, M., and Ritschel, T. (2012). Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. J. Chem. Inf. Model. *52*, 2031–2043.

28. Li, Z., Jiang, M., Wang, S., and Zhang, S. (2022). Deep learning methods for molecular representation and property prediction. Drug Discov. Today *27*, 103373.

29. 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. (2020). Machine learning in drug discovery and development Part 1: a primer. CPT Pharmacometrics Syst. Pharmacol. *9*, 129–142.

30. Dara, S., Dhamercherla, S., Jadav, S.S., Babu, C.M., and Ahsan, M.J. (2022). Machine learning in drug discovery: a review. Artif. Intell. Rev. *55*, 1947–1999.

31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Vanderplas, J. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., and Desmaison, A. (2019). PyTorch: an imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, *32* (Curran Associates, Inc.), pp. 8024–8035.

33. Hammann, F., Gutmann, H., Vogt, N., Helma, C., and Drewe, J. (2010). Prediction of adverse drug reactions using decision tree modeling. Clin. Pharmacol. Ther. *88*, 52–59.

34. Schober, P., and Vetter, T.R. (2021). Logistic regression in medical research. Anesth. Analg. *132*, 365–366.

35. Maltarollo, V.G., Kronenberger, T., Espinoza, G.Z., Oliveira, P.R., and Honorio, K.M. (2019). Advances with support vector machines for novel drug discovery. Expet Opin. Drug Discov. *14*, 23–33.

36. El-Attar, N.E., Hassan, M.K., Alghamdi, O.A., and Awad, W.A. (2020). Deep learning model for classification and bioactivity prediction of essential oil-producing plants from Egypt. Sci. Rep. *10*, 21349.

37. Gupta, A., Müller, A.T., Huisman, B.J.H., Fuchs, J.A., Schneider, P., and Schneider, G. (2018). Generative recurrent networks for de novo drug design. Mol. Inform. *37*, 1880141.

38. Blanchard, A.E., Stanley, C., and Bhowmik, D. (2021). Using GANs with adaptive training data to search for new molecules. J. Cheminf. *13*, 14.

39. Shen, M., Xiao, Y., Golbraikh, A., Gombar, V.K., and Tropsha, A. (2003). Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. J. Med. Chem. *46*, 3013–3020.

40. Manelfi, C., Gemei, M., Talarico, C., Cerchia, C., Fava, A., Lunghini, F., and Beccari, A.R. (2021). "Molecular Anatomy": a new multi-dimensional hierarchical scaffold analysis tool. J. Cheminf. *13*, 54.

41. Yoo, C., and Shahlaei, M. (2018). The applications of PCA in QSAR studies: a case study on CCR5 antagonists. Chem. Biol. Drug Des. *91*, 137–152.

42. Karlov, D.S., Sosnin, S., Tetko, I.V., and Fedorov, M.V. (2019). Chemical space exploration guided by deep neural networks. RSC Adv. *9*, 5151–5157.

43. Yasonik, J. (2020). Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. J. Cheminf. *12*, 14.

44. Guengerich, F.P. (2011). Mechanisms of drug toxicity and relevance to pharmaceutical development. Drug Metabol. Pharmacokinet. *26*, 3–14.

45. Basile, A.O., Yahi, A., and Tatonetti, N.P. (2019). Artificial intelligence for drug toxicity and safety. Trends Pharmacol. Sci. *40*, 624–635.

46. Raies, A.B., and Bajic, V.B. (2016). In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip. Rev. Comput. Mol. Sci. *6*, 147–172.

47. Rim, K.T. (2020). In silico prediction of toxicity and its applications for chemicals at work. Toxicol. Environ. Health Sci. *12*, 191–202.

48. Yang, H., Sun, L., Li, W., Liu, G., and Tang, Y. (2018). In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front. Chem. *6*, 30.

49. Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning citation. Front. Environ. Sci. *3*.

50. Aguero-Chapin, G., Galpert-Canizares, D., Dominguez-Perez, D., Marrero-Ponce, Y., Perez-Machado, G., Teijeira, M., and Antunes, A. (2022). Emerging computational approaches for antimicrobial peptide discovery. Antibiotics *11*, 936. https://doi.org/10.3390/antibiotics11070936.

51. Covell, D.G., Huang, R., and Wallqvist, A. (2007). Anticancer medicines in development: assessment of bioactivity profiles within the National Cancer Institute anticancer screening data. Mol. Cancer Therapeut. *6*, 2261–2270.

52. Huang, R., Xu, M., Zhu, H., Chen, C.Z., Zhu, W., Lee, E.M., He, S., Zhang, L., Zhao, J., Shamim, K., et al. (2021). Biological activity-based modeling identifies antiviral leads against SARS-CoV-2. Nat. Biotechnol. *39*, 747–753.

53. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *181*, 475–483.

54. Panapitiya, G., Girard, M., Hollas, A., Sepulveda, J., Murugesan, V., Wang, W., and Saldanha, E. (2022). Evaluation of deep learning architectures for aqueous solubility prediction. ACS Omega *7*, 15695–15710.

55. Ye, Z., and Ouyang, D. (2021). Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. J. Cheminf. *13*, 98.

56. DiMasi, J.A., Grabowski, H.G., and Hansen, R.W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. J. Health Econ. *47*, 20–33.

57. Mouchlis, V.D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A.G., Aidinis, V., Lynch, I., Greco, D., and Melagraki, G. (2021). Advances in de Novo drug design: from conventional to machine learning methods. Int. J. Mol. Sci. *22*, 1676.

58. Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Sci. Adv. *4*, eaap7885.

59. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. *4*, 268–276.

60. Mercado, R., Rastemo, T., Lindelöf, E., Klambauer, G., Engkvist, O., Chen, H., and Jannik Bjerrum, E. (2021). Graph networks for molecular design. Mach. Learn, Sci. Technol. *2*, 025023.

61. Li, Y., Zhang, L., and Liu, Z. (2018). Multi-objective de novo drug design with conditional graph generative model. J. Cheminf. *10*, 33.

62. Wang, M., Wang, Z., Sun, H., Wang, J., Shen, C., Weng, G., Chai, X., Li, H., Cao, D., and Hou, T. (2022). Deep learning approaches for de novo drug design: an overview. Curr. Opin. Struct. Biol. *72*, 135–144.

63. Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature *588*, 203–204.

64. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

65. Pakhrin, S.C., Shrestha, B., Adhikari, B., and Kc, D.B. (2021). Deep learning-based advances in protein structure prediction. Int. J. Mol. Sci. *22*, 5553.

66. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710.

67. Nag, S., Baidya, A.T.K., Mandal, A., Mathew, A.T., Das, B., Devi, B., and Kumar, R. (2022). Deep learning tools for advancing drug discovery and development. 3 Biotech *12*, 110.

68. Husain, A., Begum, N.A., Kobayashi, M., and Honjo, T. (2020). Native Co-immunoprecipitation assay to identify interacting partners of chromatin-associated proteins in mammalian cells. Bio. Protoc. *10*, e3837.

69. Nixon, A.E., Sexton, D.J., and Ladner, R.C. (2014). Drugs derived from phage display: from candidate identification to clinical practice. mAbs *6*, 73–85.

70. Hamdi, A., and Colas, P. (2012). Yeast two-hybrid methods and their applications in drug discovery. Trends Pharmacol. Sci. *33*, 109–118.

71. Bagherian, M., Sabeti, E., Wang, K., Sartor, M.A., Nikolovska-Coleska, Z., and Najarian, K. (2021). Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. Briefings Bioinf. *22*, 247–269.

72. Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. Bioinformatics *34*, i821–i829.

73. Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., and Lu, H. (2017). Deep-learning-based drug-target interaction prediction. J. Proteome Res. *16*, 1401–1409.

74. Lee, I., Keum, J., and Nam, H. (2019). DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput. Biol. *15*, e1007129.

75. Zhu, J., Wang, J., Wang, X., Gao, M., Guo, B., Gao, M., Liu, J., Yu, Y., Wang, L., Kong, W., et al. (2021). Prediction of drug efficacy from transcriptional profiles with deep learning. Nat. Biotechnol. *39*, 1444–1452.

76. Öztürk, H., Ozkirimli, E., and Arzucan, Özgür (2019). WideDTA: prediction of drug-target binding affinity. Preprint at arXiv. https://doi.org/10.48550/arXiv.1902.04166.

77. Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., and Aittokallio, T. (2015). Toward more realistic drug-target interaction predictions. Briefings Bioinf. *16*, 325–337.

78. He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. (2017). SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J. Cheminf. *9*, 24.

79. Karimi, M., Wu, D., Wang, Z., and Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics *35*, 3329–3338.

80. Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. Nat. Biotechnol. *30*, 679–692.

81. Murphy, E.M., Jimenez, H.R., and Smith, S.M. (2008). Current clinical treatments of AIDS. Adv. Pharmacol. *56*, 27–73.

82. Tamma, P.D., Cosgrove, S.E., and Maragakis, L.L. (2012). Combination therapy for treatment of infections with gram-negative bacteria. Clin. Microbiol. Rev. *25*, 450–470.

83. Li, P., Huang, C., Fu, Y., Wang, J., Wu, Z., Ru, J., Zheng, C., Guo, Z., Chen, X., Zhou, W., et al. (2015). Large-scale exploration and analysis of drug combinations. Bioinformatics *31*, 2007–2016.

84. Wildenhain, J., Spitzer, M., Dolma, S., Jarvik, N., White, R., Roy, M., Griffiths, E., Bellows, D.S., Wright, G.D., and Tyers, M. (2015). Prediction of synergism from chemical-genetic interactions by machine learning. Cell Syst. *1*, 383–395.

85. Preuer, K., Lewis, R.P.I., Hochreiter, S., Bender, A., Bulusu, K.C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics *34*, 1538–1546.

86. Wagner, V., Dullaart, A., Bock, A.K., and Zweck, A. (2006). The emerging nanomedicine landscape. Nat. Biotechnol. *24*, 1211–1217.

87. Shi, J., Kantoff, P.W., Wooster, R., and Farokhzad, O.C. (2017). Cancer nanomedicine: progress, challenges and opportunities. Nat. Rev. Cancer *17*, 20–37.

88. Roy, U., Rodríguez, J., Barber, P., das Neves, J., Sarmento, B., and Nair, M. (2015). The potential of HIV-1 nanotherapeutics: from in vitro studies to clinical trials. Nanomedicine *10*, 3597–3609.

89. Li, Y., Abbaspour, M.R., Grootendorst, P.V., Rauth, A.M., and Wu, X.Y. (2015). Optimization of controlled release nanoparticle formulation of verapamil hydrochloride using artificial neural networks with genetic algorithm and response surface methodology. Eur. J. Pharm. Biopharm. *94*, 170–179.

90. Muñiz Castro, B., Elbadawi, M., Ong, J.J., Pollard, T., Song, Z., Gaisford, S., Pérez, G., Basit, A.W., Cabalar, P., and Goyanes, A. (2021). Machine learning predicts 3D printing performance of over 900 drug delivery systems. J. Contr. Release *337*, 530–545.

91. Alafeef, M., Srivastava, I., and Pan, D. (2020). Machine learning for precision breast cancer diagnosis and prediction of the nanoparticle cellular internalization. ACS Sens. *5*, 1689–1698.

92. Moumné, L., Marie, A.C., and Crouvezier, N. (2022). Oligonucleotide therapeutics: from discovery and development to patentability. Pharmaceutics *14*, 260.

93. Chiba, S., Lim, K.R.Q., Sheri, N., Anwar, S., Erkut, E., Shah, M.N.A., Aslesh, T., Woo, S., Sheikh, O., Maruyama, R., et al. (2021). eSkip-Finder: a machine learning-based web application and database to identify the optimal sequences of antisense oligonucleotides for exon skipping. Nucleic Acids Res. *49*, W193–W198.

94. Dar, S.A., Gupta, A.K., Thakur, A., and Kumar, M. (2016). SMEpred workbench: a web server for predicting efficacy of chemicallymodified siRNAs. RNA Biol. *13*, 1144–1151.

95. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

96. Chen, P., Dong, W., Wang, J., Lu, X., Kaymak, U., and Huang, Z. (2020). Interpretable clinical prediction via attention-based neural network. BMC Med. Inf. Decis. Making *20*, 131.

97. Tang, Q., Nie, F., Zhao, Q., and Chen, W. (2022). A merged molecular representation deep learning method for blood-brain barrier permeability prediction. Briefings Bioinf. *23*, bbac357.