Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Smart Process Manufacturing toward Carbon Neutrality—Review

Artificial Intelligence in Pharmaceutical Sciences

Mingkun Lu [a,c], Jiayi Yin [a], Qi Zhu [a], Gaole Lin [a], Minjie Mou [a], Fuyao Liu [a], Ziqi Pan [a], Nanxin You [a], Xichen Lian [a], Fengcheng Li [a], Hongning Zhang [a], Lingyan Zheng [a,c], Wei Zhang [a], Hanyu Zhang [a], Zihao Shen [b,d], Zhen Gu [a], Honglin Li [b,d,e,*], Feng Zhu [a,c,*]

[a] College of Pharmaceutical Sciences & The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China
[b] Shanghai Key Laboratory of New Drug Design, East China University of Science and Technology, Shanghai 200237, China
[c] Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba–Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China
[d] Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200062, China
[e] Lingang Laboratory, Shanghai 200031, China

ARTICLE INFO

Article history:
Received 30 September 2022
Revised 11 December 2022
Accepted 6 January 2023
Available online xxxx

Keywords:
Artificial intelligence
Machine learning
Deep learning
Target identification
Target discovery
Drug design
Drug discovery

ABSTRACT

Drug discovery and development affects various aspects of human health and dramatically impacts the pharmaceutical market. However, investments in a new drug often go unrewarded due to the long and complex process of drug research and development (R&D). With the advancement of experimental technology and computer hardware, artificial intelligence (AI) has recently emerged as a leading tool in analyzing abundant and high-dimensional data. Explosive growth in the size of biomedical data provides advantages in applying AI in all stages of drug R&D. Driven by big data in biomedicine, AI has led to a revolution in drug R&D, due to its ability to discover new drugs more efficiently and at lower cost. This review begins with a brief overview of common AI models in the field of drug discovery; then, it summarizes and discusses in depth their specific applications in various stages of drug R&D, such as target discovery, drug discovery and design, preclinical research, automated drug synthesis, and influences in the pharmaceutical market. Finally, the major limitations of AI in drug R&D are fully discussed and possible solutions are proposed.

1. Introduction

In the past few decades, the pharmaceutical industry has been limited by the extent of cutting-edge research in pharmaceutical sciences, because the development of new drugs is a long and complex process accompanied by high risks and high costs [1,2]. In other words, the current field of drug research and development (R&D) requires significant productivity improvements to shorten the cycle time and cost of drug development [3]. Technologies such as network pharmacology, RNA-sequencing (RNA-seq), high-throughput screening (HTS), or virtual screening (VS) have all accelerated the discovery of new targets, as well as new drugs to some extent [4–9]. Nevertheless, these technologies have rarely been significant contributors to the current process of new drug discovery. Thus, there is an urgent need for new technology to drive the development of new drugs.

As the computing power of devices grows, artificial intelligence (AI) has been used in many real cases, such as in image classification and speech recognition, due to its ability to learn, process, and predict massive amounts of information [10–12]. At present, after a long period of data accumulation, in combination with the development of high-throughput RNA-seq technology, massive amounts of biomedical data have been collected [13–18]. Biomedical data, which has a high level of heterogeneity and complexity, comes from a variety of sources, including omics data from different platforms, experimental data from biological or chemical laboratories, data generated by pharmaceutical companies, publicly disclosed textual information, and manually collated data from publicly available databases [19–22]. AI can be used to learn the potential patterns in these vast amounts of biomedical data, thereby bringing new opportunities and challenges to the pharmaceutical sciences and industries.

* Corresponding authors.
E-mail addresses: hlli@ecust.edu.cn (H. Li), zhufeng@zju.edu.cn (F. Zhu).

https://doi.org/10.1016/j.eng.2023.01.014

Please cite this article as: M. Lu, J. Yin, Q. Zhu et al., Artificial Intelligence in Pharmaceutical Sciences, Engineering, https://doi.org/10.1016/j.eng.2023.01.014

The AlphaFold2 system used AI in the 14 round of the Critical Assessment of Protein Structure Prediction (CASP14) competition and outperformed others in accurately predicting the three-dimensional (3D) structures of proteins [23]. Similarly, in the Open-Graph Benchmark Large-Scale Challenge (OGB-LSC) competition, a graph neural network (GNN) combined with a transformer model won the top rank in predicting the molecular properties calculated by means of density functional theory (DFT), which is difficult and highly time-consuming using traditional methods [24]. These competitions demonstrated the strong ability of AI to analyze biological or chemical data. Due to its powerful capability to utilize related biomedical data to understand complex biological systems and chemical reaction spaces [25,26], AI has had a revolutionary impact on all stages of drug R&D, including not only research on proteins and small molecules but also the assisted design of clinical trials and post-market surveillance [27]. Furthermore, in pharmaceutical companies, many state-of-the-art (SOTA) AI models have been adopted in diverse pipelines to shorten the R&D cycle time and decrease costs [28–30].

AI techniques in this context mainly involve machine learning (ML) and deep learning (DL). Both ML and DL algorithms are involved in target discovery and validation [31], drug discovery and design [32], and preclinical drug research [33], where they are used to analyze different data characteristics in different formats. After a drug candidate is enrolled in a clinical trial [34], DL plays a pivotal role in assisting in the design of the clinical trial and in supervising and analyzing data from the clinical phase IV [33]. Approved drugs have a strong impact on manufacturing [35] and the market economy, and DL can play a part in these areas as well. Therefore, in this review, we present a comprehensive overview of most aspects of the use of AI in the pharmaceutical sciences. We focus on how AI can be used to promote target discovery and drug discovery (as shown in Fig. 1) and reflect on how to further accelerate the development of this field.

2. Basic concepts of AI and its scope of application

AI was first proposed at the Dartmouth Conference in 1956 and was defined as an algorithm that gives machines the ability to reason and perform functions [36]. From perceptual machines to support vector machines (SVMs) and artificial neural networks (ANNs), the development of AI has gone through several ups and downs, and is currently flourishing thanks to the hardware support that is now available. Both ML and DL fall under the category of AI; strictly speaking, DL can be placed within the category of ML. However, our discussion of ML in this review only concentrates on traditional ML methods, such as random forest (RF) and SVMs.

2.1. The big data era

In the current big data era, gigantic amounts of biological and clinical data have laid a foundation for the application of AI in the field of medical and pharmaceutical research. Although AI has been successfully and effectively applied in multiple aspects of the drug R&D process, the quantity and quality of medical data have become one of the main obstacles to the development of AI in the pharmaceutical sciences. Thus far, pharmaceutical databases with detailed and structured big data proposed by medicinal researchers worldwide are playing a key role in promoting AI applications in medical and pharmaceutical research.

For example, the Therapeutic Target Database (TTD) includes the most comprehensive information about known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information, and the corresponding drugs directed at each of these targets. It provides detailed knowledge of the functions of targets, as well as their sequence, 3D structures, ligand-binding properties, relevant enzymes, and corresponding drug information [37]. PubChem [17] provides collective information of chemical molecules and their activities in response to biological assays, including molecular structure, identifiers, physicochemical properties, patent information, and molecular toxicity. Some popular databases aimed at various pharmaceutical issues have been proposed and are frequently used; these play significant roles in promoting the application of AI in medical and pharmaceutical research [38–42]. Summarizing various popular pharmaceutical databases, Table 1 [17,18,37,43–62] provides brief information on popular pharmaceutical databases, categorized into protein-related, gene-related, drug-related, and disease-related databases.

2.2. ML and DL

Unlike traditional computer programming calculations, ML and DL can learn potential patterns from the input data without explicit programming. They are not limited by the format of the input data, which is broad and can include text, images, sound, and more (all types of data that can be encoded) [63]. Similar to the human learning model, ML and DL can gradually recognize different features of the data, infer the patterns lying within, and update their model parameters through continuous iterations until a valid model is formed.

According to the application scenarios, the models can be categorized into regression models and classification models. The difference between classification and regression tasks lies mainly in whether the type of output variable is continuous or discrete. Cheng and Ng [64] applied ML approaches to predict the biological activity of per- and polyfluorinated alkyl substances (PFAS) with an output of continuous values, and this study is a typical regression task. Hong et al. [65] built a DL model to predict whether a protein in a bacterium is of the type IV secreted effectors (T4SE), with an output of discrete values (e.g., 0/1), and this study is a typical classification task.

Depending on the type of learning algorithm required to solve the problem, models are conceptualized into three categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a labeled-data-driven process that trains a model on the relationship between input and its prespecified output in order to predict the categories or continuous variables of future input. In comparison, unsupervised methods are used for identifying patterns in unlabeled datasets and exploring a dataset's potential structures to allow clustering of the data for further analysis. In addition, semi-supervised learning is partway between supervised and unsupervised learning; it accepts only part of the labeled data to develop a training model and is used as a potential solution for problems that lack high-quality data [66]. Reinforcement learning performs model construction through constant interactive learning, relying on penalties for failure or rewards for success.

2.3. Introduction to different types of ML/DL-based algorithms

ML and DL methods have been successfully applied to solve relevant biomedical problems, with the adopted modeling approach varying for different problems or even the same problems. For example, small molecules used to be characterized as engineered features for direct loading in several ML methods to predict the properties; however, more recently, GNNs can also be utilized to describe small molecules for predictions of properties [67]. Determining the function annotations of proteins is essential for the selection of druggable proteins as potential targets. Kulmanov et al. [68] conducted a convolutional neural network (CNN) to annotate the gene ontology annotation (GOA) of proteins. Gligorijević

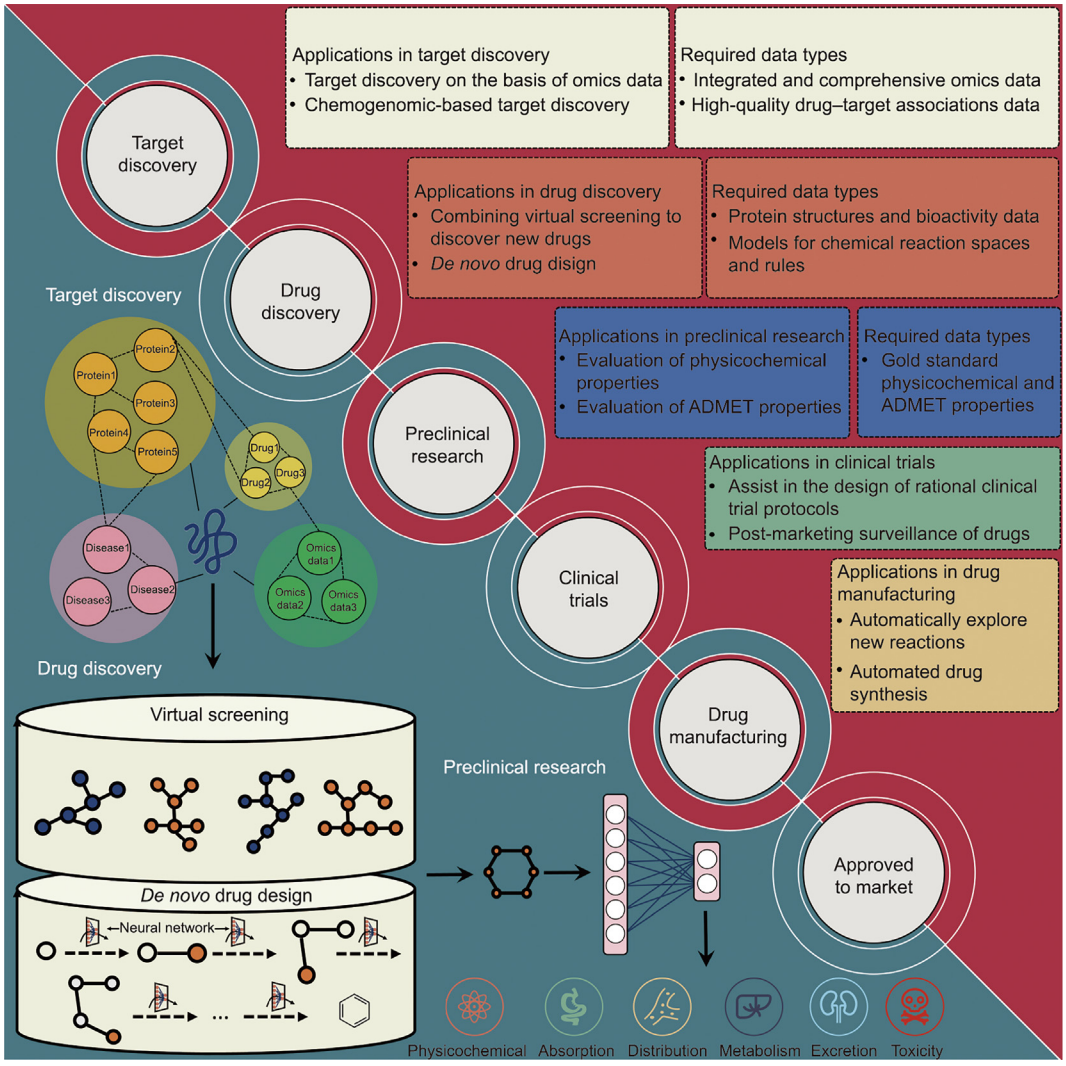

Fig. 1. Summary of AI applications in the pharmaceutical sciences. ADMET: absorption, distribution, metabolism, excretion, and toxicity.

et al. [69] built a recurrent neural network (RNN) for protein function annotations, and Xia et al. [70] combined both a CNN and RNN to predict the gene ontology (GO) label of proteins.

ML builds a special algorithm—not a specific algorithm—that focuses on the features of the data and transforms them into knowledge that machines can read to provide humans with new insights. Various common algorithms exist for researchers to choose from. The naïve Bayes (NB) algorithm is a probabilistic-based classifier based on Bayes' theorem and independence assumptions between features; it is a simple and intuitive algorithm [71]. An RF algorithm constructs a set of unrelated decision trees that form a whole hierarchical structure; under model construction, each tree is individually responsible for a corresponding problem [72]. The final decision is based on the majority votes of the decision trees. Models that make decisions based on this approach are also commonly referred to as ensemble models. Extreme gradient boosting (XGBoost) is a scalable ML algorithm based on gradient boosting, which is also an ensemble model [73]. Multi-layer perceptron (MLP) can be viewed as a directed graph consisting of multiple node layers, each fully connected to the next layer, so that it maps a set of input vectors to a set of output vectors. SVM is one of the most widely applied ML algorithms. An optimal hyperplane is used to classify samples, which are obtained by maximizing the margins between different classes in a specific dimensional space, with the dimensionality being determined by the number of features [74]. The *k*-nearest neighbor (KNN) is regarded as "lazy learning" that classifies the sample according to only a few neighboring samples when distinguishing between categories [75]. In addition to the above methods, several other ML methods such as principal component analysis (PCA), partial least-squares (PLS), linear discriminant analysis (LDA), and logistic regression (LR) have been applied in biomedical data processes [76,77].

DL is popular due to its powerful generalization and feature-extraction capabilities; its learning and prediction process is end-to-end. Unlike the traditional ML process (which often consists of multiple independent modules), DL obtains the output data (output-end) directly from the input data (input-end) during the model training process and continuously adjusts and optimizes the model based on the error between the output and the true value, until it meets the expected result. A deep neural network (DNN) is a feed-forward neural network consisting of densely connected input, hidden, and output layers. It achieves the feature learning of input data by simulating nonlinear transformations between neurons, with each layer consisting of various neurons [78]. A CNN is a feed-forward neural network that consists of con-

Table 1
Pharmaceutical databases focusing on proteins, genes, drugs/drug targets, and diseases.

Focus	Database	Description	Refs.
Proteins	RCSB PDB	PDB contains 3D structural data of large biological molecules, such as proteins and nucleic acids	[43]
	PRIDE	PRIDE is a public data repository for proteomics, including protein and peptide identifications, post-translational modifications and supporting spectral evidence	[44]
	UniProt	UniProt is a protein database containing protein sequences, functional information, and an index of research papers	[18]
	InterPro	InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites	[45]
	VARIDT	VARIDT provides comprehensive data on all aspects of drug transporters' variability	[46,47]
Genes	Ensembl	Ensembl provides centralized genomic data and powerful functionalities such as gene annotation and regulatory function predictions	[48]
	UCSC Genome	The UCSC Genome browser offers access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms	[49]
	GEO	The GEO is a database repository of high-throughput gene expression data and hybridization arrays, chips, and microarrays	[50]
	GenBank	GenBank is an annotated collection of all publicly available DNA sequences	[51]
	RefSeq	RefSeq provides separate and linked records for the genomic DNA, gene transcripts, and corresponding proteins for multiple organisms	[52]
	EA	EA collects baseline gene expression data for different species and contexts, and contains differential studies reporting expression changes under two different conditions	[53]
Drugs/drug targets	TTD	TTD includes the most comprehensive information about known and explored therapeutic protein and nucleic acid targets	[37]
	ChEMBL	ChEMBL is a manually curated library of bioactive compounds with drug-like properties	[54]
	PubChem	PubChem covers collective information on chemical molecules and their activities in response to biological assays	[17]
	DrugBank	DrugBank combines comprehensive drug target information with specific drug data	[55]
	DrugMAP	DrugMAP provides a comprehensive list of interacting molecules for drugs/drug candidates, including information on differential expression patterns	[56]
	DTC	DTC enables the exploration of bioactivity data, the processing of new bioactivity data, and data curation in order to improve the understanding of DTIs	[57]
	PHAROS	PHAROS provides a comprehensive, integrated knowledge base for the druggable genome	[58]
Diseases	TCGA	TCGA has over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data related to the cancer genome	[59]
	DisGenNET	DisGenNET contains large, publicly available collections of genes and variants associated with human diseases	[60]
	ClinVar	ClinVar is a public archive of reports on relationships among human variations and phenotypes, with supporting evidence	[61]
	OMIM	OMIM is an online catalog of human genes and genetic disorders	[62]

PDB: Protein Data Bank; PRIDE: proteomics identification database; UniProt: universal protein; UCSC: University of California at Santa Cruz; GEO: Gene Expression Omnibus; EA: expression atlas; DTC: drug target commons; DTIs: drug–target interactions; TCGA: The Cancer Genome Atlas; OMIM: online Mendelian inheritance in man.

volutional (feature extraction) and pooling (dimensionality reduction) layers. The convolutional and pooling layers help to extract all the information in a dataset without consuming too much time and computational resources [79]. An RNN is a class of ANN in which linked nodes form a directed or undirected graph along a temporal sequence. An RNN includes a feedback component that allows signals from one layer to be fed back to the previous layer. It is the only neural network with internal memory, which helps to address the difficulty of learning and storing long-term information [80]. A GNN is a connectivity model that derives the dependencies in a graph by means of information transfer between nodes in the network [81,82]. A GNN updates the state of a node according to neighbors of the node at any depth from the node; this state is able to represent the node information. The neural network architectures of the four networks described above are shown in Fig. 2.

An autoencoder (AE), which consists of an encoder and a decoder, is used to learn efficient encodings of input data. The encoding, which is generated by feeding input to the encoder, regenerates the input by the decoder. An AE is usually used for data compression and dimensionality reduction through the representation methods (i.e., the encoding) of a set of data [83]. A generative adversarial network (GAN) is composed of two underlying neural networks: a generator neural network and a discriminator neural network. The former is used to generate content, while the latter is used to discriminate the generated content [84]. Models can also be used in combination to solve a wider range of problems. For example, a graph convolution network (GCN) extends convolutional operations from traditional data (e.g., images) to graph data [85].

When a model fails to learn the underlying patterns in data features effectively and loses the ability to generalize to new data, such a problem is called model underfitting [86]. In contrast, overfitting occurs when the model is training and noise in the data fitted as a representative feature resulting in poor predictions for new data [87]. Compared with underfitting, model overfitting is more difficult to deal with. Models often become overfitted due to being overly complex or because of an underrepresentation of data. A dataset used for a model is often divided into a training set, validation set, and test set. These sets are respectively used for model training, model adjustment, and model evaluation. To put it simply, a model that works badly on both the training and test sets is an underfitted model, while a model that works well on the training set but badly on the test set is an overfitted model. Typical ways to suppress overfitting include regularization, data augmentation [88], dropout [89], early stopping, ensemble learning, and among other methods.

Researchers encountered underfitting and overfitting problems, using only one model of traditional epidemic models or ML models, when predicting the long-term trends of the coronavirus disease 2019 (COVID-19) pandemic. To address these issues, Sun et al. [90] proposed a new model called dynamic-susceptible–

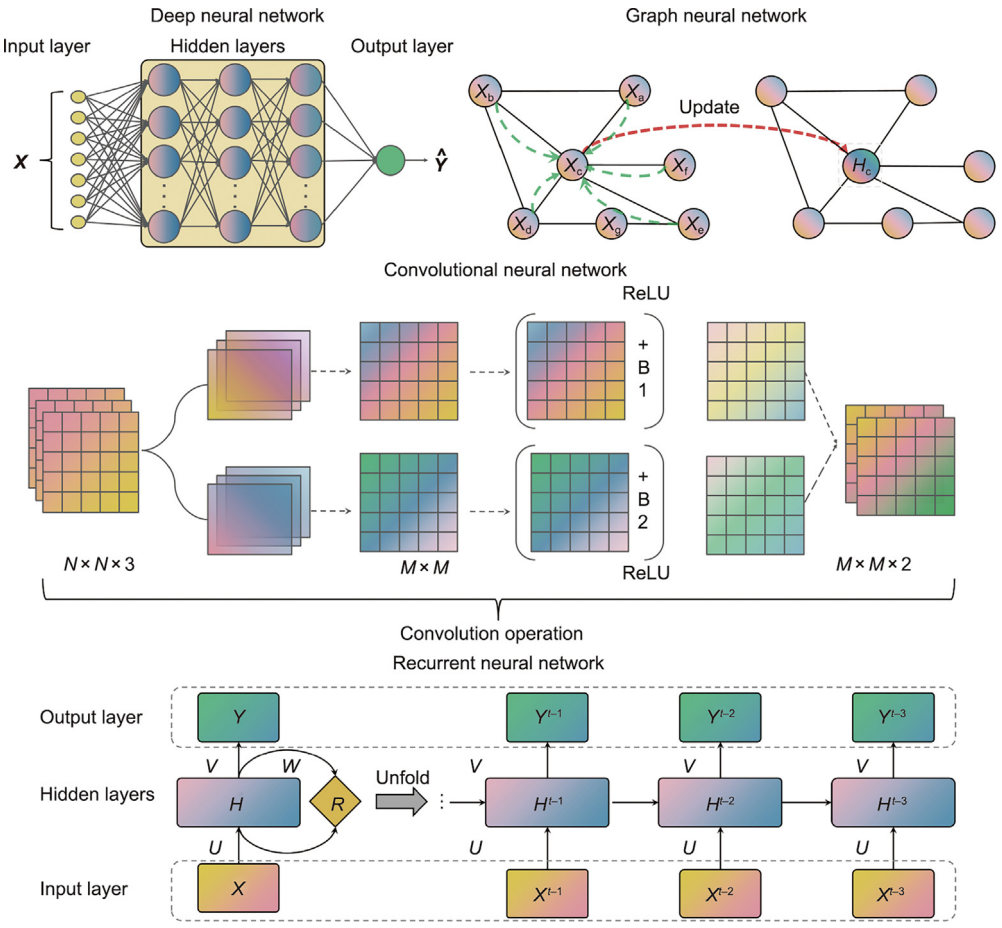

Fig. 2. Schematic network architectures for a DNN, GNN, CNN, and RNN. ReLU: rectified linear unit.

exposed–infective–quarantined (D-SEIQ). The D-SEIQ model can accurately predict the long-term trends of COVID-19 outbreaks by appropriately modifying the susceptible–exposed–infective–recovered (SEIR) model and integrating ML-based parameter optimization under reasonable epidemiology constraints.

Different models have different evaluation criteria. In regression models, commonly used evaluation criteria include mean squared error (MSE), root mean squared error (RMSE), and *R*-squared. In classification models, the more commonly used criteria are recall, precision, and F_1-score. The receiver operating characteristic (ROC) curve and precision-recall curve (PRC) are the most commonly used evaluation criteria in classification models, with ROC curves taking into account both positive and negative cases to assess the overall performance of the model, while PRCs focus more on positive cases [91].

2.4. A brief description of molecule representation as model input

Over time, the accumulation of data on small molecules and proteins has resulted in an extremely large data resource. Databases of molecular sequences, structures, physicochemical properties, and so forth have been collected and organized by different organizations and contain a great deal of knowledge and information. However, the different sources and formats of the data make it difficult to integrate the correlated data from multiple heterogeneous sources. Therefore, it is particularly important to adopt suitable methods to represent molecules in an appropriate way and to mine the crucial information in the data on molecules by means of AI [92]. Current AI algorithms are highly dependent on the quality of the data; thus, when performing model construction, it is necessary to unify the input format of molecules, such as by representing small molecules and proteins as model-readable vectors or matrices.

At present, the representation of small molecules is generally done using one of four main approaches. The first approach involves knowledge-based representation. Molecular descriptors and molecular fingerprints based on human *a priori* knowledge are widely used in various ML or DL algorithms [93]. The second approach involves direct representation based on images. CNNs have now been used to learn rules from two-dimensional (2D) digital images. A 2D chemical digital grid of a molecule can be directly used as input to allow a CNN model to learn the properties of the molecule [94]. The third approach is string-based representation. For example, a typical canonical simplified molecular-input line-entry system (SMILES) represents small molecules in the form of strings. Thus, CNNs and RNNs can be further used to learn molecular embeddings from the string representations of chemical structures [95–97]. The fourth approach involves graph-based feature representation. Representation methods based on graph convolution or graph attention have been widely used to explore the feature representation of small molecules. In these methods, atoms and bonds are considered to be nodes and edges, respectively, while new molecular representations are obtained during the continuous updating of information at individual nodes. Graph-based representations have achieved outstanding performance in a variety of pharmaceutical learning tasks [98,99].

Protein representation methods can be basically classified into four categories: representation based on intrinsic properties of

sequences, representation based on physicochemical properties, representation based on protein structure, and graph-based representation. Sequence-based protein representation methods include amino acid composition (AAC), dipeptide composition, autocorrelation descriptors, position-specific scoring matrices (PSSMs), and one-hot encoding [100–107]; these methods reflect the content of various amino acids, dipeptide content, and the distribution of amino acids on the sequence. Physicochemical property-based protein representation methods include composition, transition, and distribution (CTD), pseudo-amino acid composition (PAAC), and amphiphilic pseudo-amino acid composition (APAAC) [108–110], which reflect the properties of each amino acid and the distribution of these properties on the sequences. The two feature representation methods described above are widely used in various models, because they can obtain protein feature representations by knowing only the sequence information. It is well known that the high-level structure of a protein determines the function of that protein, so it will sometimes directly represent the structure of proteins. Protein representation methods based on structural properties include topological molecular structure and protein secondary structure and solvent accessibility (PSSSA) [111–113], which reflect the structural properties of each amino acid in a protein and the structural type of a protein. PSSSA is also a graph-based protein representation. In the simplest graph, each node corresponds to a residue, while the edges connect pairs of residues within a certain distance [114]. Structure-based and graph-based protein representation methods can effectively represent the structure of a protein and the relationships between amino acid residues in the structure, and can be applied to a variety of novel model architectures, such as GNNs, transformer models, and GANs [114–117].

In recent years, novel molecular representation methods have been emerging, such as knowledge-graph-based and large-scale pretrained-based representation methods [118,119]; these methods also excel in suitable downstream tasks. Overall, representing the raw data of a molecule using a vector or matrix that captures the molecule's key features is critical for subsequent data exploration and analysis.

2.5. The study of drug research and disease with distinct AI algorithms

When studying different types of drugs and performing disease research, choosing a suitable model can maximize the potential information of the data. Given classification or regression problems with small datasets, ML can often achieve a satisfactory performance in a short time. For example, a drug–protein affinity prediction study based on quantitative structure–activity relationship (QSAR) models could choose to use SVM or RF models (see Section 5 for more detail) [120,121]. When the amount of data is progressively higher, DL algorithms are often more appropriate. For example, for the prediction of protein-folding problems, CNN models can better predict residues [122]. In the research area of drug *de novo* design, generative models and variational autoencoders (VAEs) can help to design molecules that align with the design vision [123,124] (see Section 4 for more detail). Instead of selecting models from the perspective of the tasks, studies often use the data representation form to select an appropriate algorithm. Therefore, researchers can often choose from different AI algorithms that are available for the same task. When predicting the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of molecules, CNNs, RNNs, and multi-task learning can achieve outstanding results [125] (see Section 5 for more detail). By starting from the relationships between data, graph-based AI algorithms allow the modeling of unstructured data. In the pharmaceutical sciences, there is never a lack of complex relationships. Therefore, modeling complex interactions such as drug–drug interactions, drug–protein interactions, protein–protein interactions (PPIs), and so forth enhances the learning capability of the models [126] (see Section 3 for more detail). When combined with representations of these entities themselves, key information about the entities can be learned at a deeper level to aid in making predictions, while providing a more explanatory model.

Therefore, the boundaries between the use of distinct algorithms have become increasingly blurred when such methods are applied to the actual drugs and disease problems to be studied. Depending on the type of data available and taking into account the biological significance can be informative for model selection and construction.

3. Target identification and validation

From a conventional standpoint, there are two paradigms for discovering new (first-in-class) drugs [127]: phenotypic drug discovery (PDD) and target-based drug discovery (TDD). Early biological research techniques relied on microscopy, imaging, and cellular techniques to observe the phenotypic changes in living systems. PDD is used to screen a library of compounds or antibodies by constructing an animal model or experiment that is highly relevant to the disease. Next, the responses of cells or experimental animals to these compounds are observed, with the aim of identifying molecules with a certain level of efficacy for further structural modification and optimization [128]. With the development of molecular biology and various sequencing techniques, research on biological macromolecules has reached a new height. Drug discovery research has entered the TDD era [129], and TDD has gradually replaced PDD as the mainstream drug discovery paradigm. TDD is centered on a "one gene, one drug, and one disease" concept [4]. This approach relies on a highly disease-relevant target, which could be an enzyme, protein, or other gene product, along with an elaborate and meticulous small-molecule design for this target, which is used to modulate the target to act as a therapeutic agent for the disease. Although the drug discovery paradigm of PDD has been re-emerging in recent years [128], the screened drugs often require further target validation and mechanistic studies. Therefore, target discovery is often the first, critical step in the drug development phase [129]. The target discovery process involves multifaceted research, including the study of disease-related genes, signaling pathways, protein interactions, and small molecule–protein interactions. Of particular interest is the fact that target discovery based on experimental means is difficult to carry out quickly and widely, due to limitations in throughput, accuracy, and cost, whereas AI-based discovery can efficiently and effectively identify biomolecules with the potential to become drug targets.

3.1. Target identification based on omics techniques

With the advancement of high-throughput sequencing technologies, huge amounts of omics data are continuously being generated. The processing and analysis of such large-scale omics data (genomics, transcriptomics, proteomics, metabolomics, etc.) [130–138] have been revolutionary to biology, medicine, and pharmacology, especially in facilitating researchers' understanding of complex biological systems and processes. Many genes or proteins playing important roles in biological processes that may be associated with specific diseases have been identified based on omics data [135,139–141], thereby facilitating research on drug target discovery. For example, new candidate disease targets such as SETD2 and VGLL4 have been uncovered using omics data. However, processing and analyzing these complex and high-dimensional omics data is extremely challenging; thus, ML and DL approaches can be used to learn potential knowledge from large-scale omics datasets, which can help in the discovery of

genes or pathways critical to biological processes [142]. Table 2 [18,44,48–50,53,143–151] provides examples of omics projects for drugs, proteins, and diseases analysis.

Potential targets are molecules that are associated with a specific disease and have the smallest possible degree of association with other diseases. Complex diseases such as oncological, cardiovascular, and immune diseases are often regulated by multiple key genes, molecules, or signaling pathways, so it is often necessary to unravel the connection between multiple molecules and the disease. Omics data are essential for discovering and assessing the biological effects or toxicity of potential targets. For example, cancer stem cells (CSCs) cause great resistance to the treatment of lung adenocarcinoma (LUAD). Studying the expression of stem-cell-related genes in LUAD could provide new insights into the treatment of LUAD. Zhang et al. [152] applied an unsupervised ML algorithm known as one-class LR (OCLR) to the molecular datasets of normal stem cells and their progeny to obtain the messenger RNA (mRNA) expression-based stemness index (mRNAsi), DNA methylation-based stemness index (mDNAsi), and epigenetic regulation based mRNAsi (EREG-mRNAsi) for analyzing the LUAD cases data in The Cancer Genome Atlas (TCGA) in order to calculate the scores of sample stemness indices. In this process, weighted gene co-expression network analysis (WGCNA) was used to find key genes associated with LUAD. In the end, 13 previously overlooked key genes with an overall association were identified, which could be used as potential targets for the treatment of LUAD by suppressing the stemness features.

Since their release, the connectivity map (CMAP) and Library of Integrated Network-based Cellular Signature (LINCS)-L1000 databases—which contain a large amount of transcriptomic data following drug perturbations and various other environmental disturbances—have been used to do a great deal of research to identify the mechanism of action and targets of small molecule compounds, with the aim of discovering potential drugs for diseases or potential targets for drugs [153–155]. The web service PharmMapper [156–158] gathered 52 431 pharmacophore models from TargetBank, DrugBank, BindingDB, and the potential drug target database (PDTD), and used them to identify potential target candidates for the given probe small molecules by means of a fast pharmacophore mapping approach. ChemMapper [159] is another web service that aims to predict polypharmacology effects, potential protein targets, and modes of action for small molecules based on 3D similarity computation, using a database containing 4 350 000 chemical structures with bioactivities and associated target annotations. The iDrug [160] platform provides a versatile, user-friendly, and efficient online tool for computer-aided drug design (CADD) based on pharmacophore and 3D molecular similarity searching, enabling binding sites detection, VS, and drug target prediction in an interactive manner through a seamless interface. DeltaNet was designed by Noh and Gunawan [161] based on the ordinary differential equation (ODE) model for analyzing gene transcription processes and predicting potential targets of compounds. There are two versions of DeltaNet—namely, DeltaNet-LAR and DeltaNet-LASSO—which use last angle regression (LAR) and least absolute shrinkage and selection operator (LASSO) regularization to solve linear regression problems, respectively. DeltaNet outputs a predicted ranking of gene targets for further enrichment analysis to find other key molecular targets. Zhu et al. [162] constructed a DL-based efficacy prediction system (DLEPS) to identify new drug candidates and discovering targets. Trained by transcriptional profiles data, mainly from the L1000 project profiles, DLEPS uses changes in gene expression profiles in the state of disease as input. In addition to the discovery of three new drug candidates, DLEPS also demonstrated that mitogen-activated protein kinase kinase (MEK)–extracellular-signal-regulated kinase (ERK) was a critical signaling pathway in nonalcoholic steatohepatitis—knowledge that can be used to develop specific targets. The data mining analysis of such transcriptomes through ML and DL can help not only to find drug targets but also to elucidate the mode of action of drugs and disease mechanisms [163].

The analysis of omics data has helped researchers to identify many overlooked disease candidate targets [164]. With the advancement of sequencing technology and deeper research, the drawbacks of the deeper mining of only single omics data are becoming increasingly obvious, as such mining can neither reflect the relevance and variability of biological processes (e.g., simple gene expression levels do not reflect true protein expression levels) nor reveal complex biological systems and disease mechanisms (e.g., glycolytic processes are associated with genomics, proteomics, and metabolomics). In particular, disease onset often involves multiple pathways and requires the integration of multimodal data. For example, genes with increased DNA copy numbers have been found to be involved in important cancer pathways, and somatic mutation frequency and expression levels are also important factors in cancer drivers [143,165,166]. By integrating information at multiple omics levels and mining the linear or nonlinear associations through AI approaches, candidate key factors can be identified at a more in-depth level, which is crucial for discovering candidate targets for diseases.

Complex diseases such as cardiovascular disease, schizophrenia, cancer, and Alzheimer's disease (AD) have many therapeutic targets, and multiple potential causative genes can be discovered through the multi-omics features of individual patients. Jeon et al. [31] used an SVM algorithm with a radial basis function (RBF) kernel to construct three models to predict potential targets specific to breast cancer (BrCa), pancreatic cancer (PaCa), and ovarian cancer (OvCa), respectively. Gene essentiality, gene expression, DNA copy number variation, somatic mutation, and PPI network topology were the main input features, and the SVM was able to deeply explore the association of and difference among these features to distinguish potential drug targets from non-target proteins. The model was cross-validated with ten folds and had a high area under the ROC curve (AUROC) value and a low false-positive rate. By using the trained model to predict 15 663 human proteins and score the prediction results, a total of 122 global cancer targets were identified for all cancers (69 of which corresponded to the 116 known targets that were rigorously validated). In addition, a large number of potential targets specific to BrCa, PaCa, and OvCa were identified. Of course, the identified targets were only for guidance and were not true drug targets.

Moreover, using multi-omics data with PPI networks, a group developed a network-based Bayesian algorithm framework [167] to infer loci for an AD genome-wide association study (GWAS) and revealed 103 AD risk genes (ARGs). This study included gene expression data from single cell transcriptomics, gene expression data from microarrays, and proteomics, fully demonstrating the ability of AI approaches to integrate multi-source and multimodal data to discover potential therapeutic targets.

ML has been instrumental in driving the learning process of multi-omics data, but it can be overwhelmed by larger multi-omics data and more complex problems. However, DL can handle much larger amounts of multi-omics data and unearth deeper associations. On the assumption that the drug inhibition of targets and target gene knockdown (KD) should lead to the occurrence of similar biological processes, resulting in similar mRNA expression profiles, Pabon et al. [168] explored the direct feature correlation and indirect feature correlation between compound-induced features and gene KD in CMAP, and combined these features with other features such as PPIs as inputs into the RF model to predict drug targets. To better mine the correlation between chemical perturbation (CP) features and KD genetic perturbation features,

Table 2
Omics projects for analysis of drugs, proteins, and diseases.

Omics	Project	Introduction	Data scale	Applications for drug/target discovery	Ref.
Genomics	Ensembl	Ensembl is a genome browser that can perform gene annotation, multiple alignment calculations, regulatory function predictions, and disease data gathering	622 461 regulatory features, 118 epigenomes	Mining disease genes; disease risk prediction	[48]
	UCSC Genome	The UCSC Genome browser contains information on inter-genomic alignment, different sequences, phenotypes, expression profiles, regulatory information, conservations, variants, repetitive regions, and other information	N/A	Mining disease genes; disease risk prediction	[49]
	TCGA	TCGA collects and functional genomics data including mutation, copy number, mRNA and protein expression	21 773 genes, 2 730 388 mutations	Discover novel molecular targets	[143]
Transcriptomics	GEO	The GEO is a database repository of high-throughput gene expression data and hybridization arrays, chips, and microarrays	4 348 datasets, 182 700 assays	Retrieve drug, gene, and disease perturbations	[50]
	EA	EA collects baseline gene expression data in different species and contexts. It also contains differential studies, reporting changes in expression between two different conditions	4 315 studies, 153 212 assays	Drug discovery and validation; disease genes analysis	[53]
	L1000	The LINCS-L1000 data repository generates gene expression signatures by treating various cell types with perturbagens, which include a variety of small-molecule compounds, gene overexpression and gene KD agents	~1 300 000 profiles	Drug/target discovery; drug repositioning	[144]
Proteomics	PRIDE	The PRIDE is a public data repository for proteomics, including protein and peptide identifications, post-translational modifications, and supporting spectral evidence	19 990 datasets, 152 961 704 protein evidence	Drug target identification	[44]
	UniProt	UniProt is a protein database containing protein sequences, functional information, and indexing of research papers	568 002 proteins	Druggable proteome analysis; drug target identification	[18]
	HPA	The HPA portal is a publicly accessible database that contains millions of high-resolution images illustrating the spatial distribution of proteins in 46 different human cell lines, 20 different cancer types, and 44 different normal human tissues	27 173 antibodies targeting, 17 268 unique proteins	Druggable proteome; drug target efficacy and specificity	[145]
	HPM	The HPM portal is an interactive resource that incorporates the vast amount of peptide sequencing data from the human proteome project's draft map	17 294 genes, 30 057 proteins	Drug target identification; biomarkers identification	[146]
Metabolomics	HMDB	The HMDB is an open-access electronic database that provides comprehensive information on the small-molecule metabolites that have been discovered (and experimentally confirmed) in the human body	50 336 pathways, 18 198 reactions, 251 986 metabolites	Drug metabolism analysis	[147]
	Metlin	Metlin is a powerful metabolite identification and information database that contains categories of lipids, amino acids, carbohydrates, toxins, small peptides, and natural products	~860 000 molecules	Drug metabolism analysis	[148]
	KEGG pathway	KEGG pathway is a collection of manually drawn pathway maps with a primary focus on metabolic pathways and the integration of metabolic, gene, and protein pathway information	552 pathways	Drug metabolism; drug development; disease/drug information	[149]
	MetaCyc	Metabolic pathways from all domains of life (MetaCyc) contains the pathways for primary and secondary metabolism, as well as the metabolites, reactions, enzymes, and genes that go along with them	2 937 pathways, 17 780 reactions, 18 124 metabolites	Pathway-based target selection and validation	[150]
	Reactome	Reactome is an open-source pathway database that provides user-friendly bioinformatics tools for the visualization, interpretation, and analysis of pathway knowledge	2 585 pathways, 14 246 reactions, 11 291 proteins	Simulate impact of drugs on pathway activities	[151]

mRNA: messenger RNA; LINCS: Library of Integrated Network-based Cellular Signature; KD: knockdown; HPA: human protein atlas; HPM: human proteome map; HMDB: human metabolome database; KEGG: Kyoto Encyclopedia of Genes and Genomes; N/A: not available.

Zhong et al. [169] proposed a GCN model known as Siamese spectral-based graph convolution network (SSGCN) to mine transcriptomic data to predict compound–protein interactions (CPIs). SSGCN constructed two parallel GCN models for the feature extraction of CP profiles and KD profiles, respectively, where CP profiles and KD profiles were integrated with a PPI network (the attribute values of the network nodes were gene differential expression values, and if there was an interaction between two nodes, these two nodes were connected by an edge). Two sets of graph embedding vectors were obtained after feature extraction, and the degree of correlation between the CP features and KD features was obtained by means of a simple linear regression layer. The correlation was expressed as Pearson's coefficient R^2 and was fed to the classifier as features along with cell line, CP time, dosage, and KD time to discriminate the interaction of compounds with the corresponding proteins. This model was subsequently validated externally and shown to be effective in identifying potential drug targets and facilitating drug repositioning studies.

Most of these target discovery models use end-to-end models to directly discover druggable proteins. DL can also perform key roles in multiple specific steps in the target discovery process, such as predicting splicing from pre-mRNA transcript sequence using SpliceAI [170], using scVI to predict and analyze gene expression probabilities in single cells from transcriptomic data [171], and using PLEDA to predict an enhancer predictor [172]. Some studies have performed a GWAS of COVID-19, with results suggesting a possible association with COVID-19 susceptibility in the 3p21.21 region of the chromosome. Building on these studies, Downes et al. [173] used multiple DL approaches combined with multi-omics data to discover that the gain-of-function risk A allele of a single-nucleotide polymorphism (SNP), rs17713054G>A, may be a variant that can cause disease. Further analysis revealed that leucine zipper transcription factor like 1 (*LZTFL1*), a gene regulated by rs17713054, was a critical gene for the development of epithelial-mesenchymal transition (EMT). EMT is a developmental pathway associated with lung inflammation that is frequently induced by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in lung cancer cell lines (CCLs) and the respiratory tract. As a key gene in this series of biological processes, *LZTFL1* could serve as a potential therapeutic target.

The use of AI approaches can help effectively predict drug responses in cancer cells to advance precision medicine [174–176]. One group used elastic net regression and RF to identify how multi-omics data affect drug response prediction [177]. In this study, 265 drugs across 990 CCLs were screened to construct pharmacogenomic datasets. To comprehensively investigate the influence of different combinations of molecular data, linear and nonlinear ML models were built. Among the genome-wide gene expression, DNA methylation, gene copy number, and somatic mutation data, gene expression data was the most predictive data type in pan-cancer analysis, and genomic data (i.e., driver mutations, copy number alterations, or DNA methylation data) was the most predictive data type in cancer-specific analysis.

The importance of multi-omics data in drug response prediction has also been demonstrated. However, most methods do not take drug/cell line specificity, drug/cell line, or drug–protein associations into consideration. To address this issue, Peng et al. [178] combined multi-omics data with a GCN to construct an end-to-end model known as MOFGCN. Drug/cell line associations were used to initially construct a heterogeneous network in which the nodes were drugs or cell lines. The properties of the drugs were obtained by calculating the similarity of molecular fingerprints, and the properties of the cell lines were obtained by fusing multi-omics data (i.e., gene expression, copy number variation, and somatic mutation data) and calculating their similarity. The completely constructed heterogeneous network served as the input to a graph convolutional network, and the final features were obtained by passing messages between nodes to further learn the potential associations of drugs and cell lines. To predict drug sensitivity, a CCL–drug correlation matrix required further reconstruction based on a linear correlation matrix that was calculated from the updated features of drug and cell lines. The DL framework of predicting drug sensitivity, DeepDRK [179], integrated mutations, copy number variation, DNA methylation, gene expression, and drug screening as cell line features and extracted molecular–protein information as drug features. Then, the two features were spliced as the features of a CCL–drug pair and were fed into the DNN to predict the drug sensitivity.

The combination of omics data and AI methods can help researchers quickly obtain the information they need at the molecular scale, as the various levels of omics data reflect the various processes of life activity. Integrating and analyzing this information can aid in the understanding of complex biological systems and thus assist in the discovery of new drug targets.

3.2. Drug–target interactions (DTIs) discovered using chemogenomics

The identification of DTIs is currently contributing to research in drug discovery. Newly discovered DTIs can be used to find new targets that interact with existing drugs or to discover new compounds that interact with a disease-related target. Therefore, research results on DTIs are widely used in the fields of lead compound discovery, new target discovery, drug repositioning, and drug side-effect prediction [3,180,181]. Although HTS have been developed to determine the activity of thousands of compounds at once, they cannot catch up with AI methods in terms of either cost consumption or the number of compounds measured. In general, methods for predicting DTIs have been divided into three main approaches: ligand-based methods, structure-based methods, and chemogenomic methods. Each of these three methods has its own advantages and disadvantages, with the third method being the most widely applicable and popular. Therefore, this section focuses on reviewing chemogenomic methods, while the other two methods are covered in Section 4.

The chemogenomic approach not only uses drug-related and target-related information but also connects this information to multiple sources of biomedical information in order to better predict DTIs. Publicly accessible database resources contain a large amount of structured and unstructured biomedical data to support access to information. ML and DL can extract relevant functional information and reduce the noise from this large amount of heterogeneous data in order to discover new protein targets precisely and efficiently. Table 3 [37,54,55,57,58,182–191] lists some currently high-quality public databases.

Prediction of DTIs is usually regarded as a binary classification problem. It is very convenient to use an ML approach to predict DTIs, which usually only requires obtaining the SMILES of small molecules and the sequences of target proteins. These sequences are converted into feature vectors via different rules and are later used as inputs to a model to predict their final classification. These molecules and proteins are characterized in a variety of ways and often contain information about the physicochemical properties of the molecules and proteins, as well as their structure. A number of toolkits and libraries for molecule and protein representations have been developed and are listed in Table 4 [192–215]. For example, small molecules characterized using MACCS fingerprints were spliced with protein vectors characterized by CTD descriptors and used as inputs to an SVM to predict DTIs [216]. The occurrence of a DTI is influenced by numerous factors and corresponds to multidimensional features that represent the structure and properties of the molecule and protein. It is hoped that the model can find out more about the mechanism of DTI from these features and then

give classification judgments based on information. Such problems have also been treated as regression problems; DeepDTA is a CNN model that used the SMILES of small molecules with sequences of proteins to predict the affinity of small molecules with proteins [217]. Using only single-feature representation does not fully characterize small molecules or proteins, so some studies have used multiple descriptors to characterize small molecules and proteins and have integrated these features as vectors of inputs to predict DTIs. This improves the classification performance of the model to a certain extent [218]. In order to enable researchers to more conveniently use DL to make predictions about DTIs, Huang et al. [219] proposed DeepPurpose, which implements more than 50 DL models (including CNN, MLP, RNN, etc.). DeepPurpose can encode proteins in seven distinct ways, including MLP on AAC, PAAC, conjoint triad, quasi-sequence descriptors, CNN on amino acid sequences, RNN on top of CNN, and transformer encoder on substructure fingerprints. For compounds, there are eight encoders, including MLP on Morgan, PubChem, Daylight and RDKit 2D fingerprint, CNN on SMILES strings, RNN on top of CNN, transformer encoders on substructure fingerprints, and a message-passing GNN on a molecular graph. Those encoding methods just use SMILES and the amino acid sequence as input. In this way, researchers can conveniently predict DTIs using different encoding methods on different models.

The abovementioned studies were able to obtain a good performance using only the SMILES sequence and amino-acid sequence of proteins. At the same time, it is important to integrate various data sources to predict DTI, such as drug–drug interactions, PPIs, and drug–disease associations. Bleakley and Yamanishi [220] constructed a bipartite graph on DTI [221,222] and applied an SVM model for DTI prediction in a later work. The four datasets constructed in this work have become the gold standard datasets for later DTI prediction models. Inspired by this work, there have been a proliferation of network-based approaches to predict DTI. A computational pipeline called DTINet was then developed that integrated multiple heterogeneous data sources to construct networks on DTI [223]. In this study, four drug similarity networks were constructed based on ① drug–drug interaction networks, ② drug–disease association information, ③ drug side-effect association information, and ④ chemical structure information. Similarly, three protein similarity networks were constructed based on ① PPIs, ② protein–disease associations, and ③ genomic sequences. Using these similarity networks, a network diffusion algorithm (random walk with restart (RWR)) was first applied on individual networks separately, and the feature vectors were optimized. The low-dimensional vector representations obtained after this learning process contained information derived from various heterogeneous data sources and were able to better represent the drug/protein-specific properties. The obtained vectors were then used to discover new DTIs according to their spatial correspondence with drugs and proteins.

The use of DL models allows for the integration of heterogeneous data from multiple sources while providing a comprehensive characterization of drugs or biomolecules. Zeng et al. [224] proposed a framework called deepDTnet to integrate heterogeneous data sources for the prediction of DTI. In this study, 15 networks—including genomics, GOA, protein-related similarity, and drug-related similarity—were integrated to construct a heterogeneous network connecting drug targets and disease information. A DNN for graph representation (DNGR) algorithm was developed to obtain the informative vector of both drugs and targets based on the constructed network. However, the lack of negative samples in public databases led to difficulties in the model training process; thus, a PU-matrix completion algorithm was employed to infer whether two drugs shared a target. The results showed that combining the heterogeneous data to re-represent the drug and target without a descriptor or fingerprint achieved an excellent performance.

As mentioned before, the emergence of large-scale knowledge of omics data, systems biology, chemistry, pharmacology, and so forth is providing new perspectives for DTI prediction. However, the integration of heterogeneous data from multiple sources undoubtedly introduces a huge amount of noise and does not solve the "cold-start" problem well. Here, knowledge graphs (KG) stand out with their powerful ability to integrate heterogeneous information. By leveraging the interactions of phenotype, drug, target, and gene, a KG can help to further understand the molecular mechanism of a disease and to explore potential drug targets. Recent studies have integrated resources from several databases (DrugBank, TTD, ChEMBL, BindingDB, SIDER, GO, etc.) to construct KG such as BioKG, PharmKG, Hetionet, and Drug Repurposing (DRKG) [30,225]. A KG usually represents knowledge as a triple, which is composed of a head entity, relation, and tail entity. In the field of DTI recognition, the KG embedding (KGE) model is often used to represent entities and relations by means of low-rank vectors, in what is also known as the representation learning of KGs. The representation vectors obtained by a KG can be further used for link prediction to discover drug–target relationships [30]. A KG typically integrates a huge amount of data with dozens or even hundreds of relationships. The vectors obtained via a KG often contain a certain exact positioning and relationship of this entity in the biological network, but not its own structure or physical and chemical properties. The same is true for proteins. To address this issue, Ye et al. [118] developed a framework called KGE_neural factorization machine (NFM) that performs DTI prediction using a KGE technique combined with a recommendation system technique. In this process, an accurate entity vector is first obtained from the potential information learned from the heterogeneous network via KGE. Next, the structural information of the drug and target is obtained from molecular fingerprints and protein descriptors. Finally, multimodal information is extracted using aNFM, and the DTIs are predicted using DL methods. This approach was tested for "cold-start" scenarios of drugs or proteins and achieved a SOTA performance, particularly for protein "cold-start" scenarios.

In addition to the aforementioned methods for predicting DTIs, similarity-based [226] and matrix decomposition-based methods [227] can be used, among others, and have contributed greatly to DTI prediction in the past. With the development of DL, network-based methods, feature-based methods, and so forth are now being used in combination, bringing the advantages of each method into play to better predict DTIs and discover new targets [228,229]. Based on recent studies in the field, DTI research methods can be roughly classified into six groups; Table 5 [217,221,223,226,227,230–247] provides a brief summary of the relevant strategies.

Future research should integrate omics data more closely with biomedical data networks for a more accurate characterization of drugs or proteins. Moreover, similarity approaches have a crucial effect on DTI prediction, and combining multiple similarity results may improve model performance. One common problem in model training is the unavailability of accurate negative datasets. Accurate DTI data in publicly available data sources are rigorously experimentally validated, and the experimental validation process for each one is exhaustive; however, most failed experiments will not be reported. Furthermore, manually validated data is time-consuming, and a large amount of data has not been validated for exact interactions. Therefore, the dataset used for DTIs should always use the latest and most comprehensive drug–target database, such as TTD and DrugBank, and additional inactive experimental data should be open-sourced to improve the current DTI data system.

Table 3
Databases for DTIs research.

Database	Description	Focus	Items					Relations	Target scale	Drug scale	Ref.
			Protein or target	Drug or chemical	Disease	Gene	Pathway or mechanism				
TTD	TTD is a database providing information about known and explored therapeutic proteins and nucleic acid targets, along with the targeted disease, pathway information, and corresponding drugs directed at each of these targets	Drug, target	√	√	√	√	√	Drug–drug, drug–target, drug–disease	3 578	38 760	[37]
ChEMBL	ChEMBL is a manually curated library of bioactive compounds with drug-like properties. Data on chemicals, biological activity, and genomics are integrated	Drug, target	√	√	√	–	√	Drug–target, drug–disease	15 072	14 293	[54]
DrugBank	The DrugBank database combines comprehensive drug target information with specific drug data	Drug, target	√	√	√	–	√	Drug–drug, drug–target, drug–disease	4 563	14 748	[55]
DTC	DTC is a platform that enables the exploration of bioactivity data, the processing of new bioactivity data, and data curation in order to improve the understanding of DTIs	Drug, target	√	√	–	–	–	Drug–target	1 007	4 276	[57]
Pharos	Pharos is a database that provides a comprehensive, integrated knowledge base for the druggable genome in order to illuminate the part of the genome that has not been well described or annotated	Target, disease	√	√	√	√	–	Drug–target, drug–gene, drug–disease	20 412	1 737	[58]
Comparative Toxicogenomics Database	Comparative Toxicogenomics Database is a database that aims to increase understanding about how environmental exposures impact human health. It offers manually curated data on links between chemicals, genes, and proteins, as well as between diseases and chemicals	Drug, gene, disease		√	√	√	√	Drug–target, drug–gene, drug–disease	N/A	1 498	[182]
GtoPdb	GtoPdb is a database of ligand-activity–target interactions that contains quantitative data on pharmacological targets and the experimental and prescribed medications that affect them	Drug, target	√	√	√	–	–	Drug–target, drug–disease	3 002	11 348	[183]
DrugCentral	DrugCentral is an online resource for drug information that provides details on pharmaceutical products, active ingredients, chemical entities, pharmacological modes of action and indications, and pharmacologic actions	Drug, target, disease	√	√	√	–	√	Drug–target, drug–disease	N/A	4 714	[184]
TDR Targets	TDR Targets is a database that contains information on targets, drugs, and/or biologically active molecules of interest and can be used to prioritize targets across the whole genome	Target, disease	√	√	–	√	√	Drug–target, drug–gene	254 986	238 286	[185]
DockCoV2	DockCoV2 is a drug database with 3 548 compounds. It predicts the binding affinity of SARS-CoV-2 related proteins and 67 human proteins	Drug, target	√	√	–	–	–	Drug–target	7	3 109	[186]
KEGG Drug	KEGG Drug is a comprehensive drug information repository for approved drugs, containing information on therapeutic targets, drug metabolism, and other molecular interaction networks	Drug	√	√	√	√	√	Drug–drug, drug–target	N/A	11 952	[187]
PDID	The PDID contains a large number of potential and native protein–drug interactions in the structural human proteome	Drug, protein	√	√	–	–	–	Drug–target	3 746	51	[188]
SIDER	SIDER includes information on approved drugs and ADRs, drug side-effect frequency and side-effect categories, and drug–target relations	Drug, side-effect	√	√	√	–	–	Drug–side effect, drug–disease	5 868	1 430	[189]
Open Targets	The Open Targets platform is a comprehensive platform for identifying and prioritizing potential therapeutic drug targets. It generates and scores target–disease connections. It also includes annotation information on targets, diseases, phenotypes, drugs, and their interactions	Target, disease	√	√	√	–	–	Drug–target, drug–disease	61 524	12 854	[190]
PDTD	The PDTD is a protein database for target identification containing 1 186 protein entities that cover 831 known or potential drug targets. It also provides annotations including protein and active site structures, related diseases, biological functions, and associated regulating (signaling) pathways	Target, disease, biological function, pathway	√	–	√	–	√	Target–disease, target–pathway, target–biological function	1 186	N/A	[191]

GtoPdb: guide to pharmacology database; TDR: tropical disease research; PDID: protein–drug interaction database; SIDER: side-effect resource; ADR: adverse drug reaction.

4. SOTA application of AI to modern drug design

Drug discovery is a long-term and painstaking process. In the past decades, techniques such as HTS and combinatorial chemistry, as well as other techniques, played an important role in the discovery of lead compounds. Further structural modifications of the obtained lead compounds were then developed to reduce toxicities and improve efficacy. As these techniques gradually increased in popularity, however, their various disadvantages were gradually revealed. Similarly, in the 1980s, CADD was no less popular than today's AI. For example, QSAR were widely used as soon as they were proposed. However, in those days, QSAR-based models were limited by the available computing power, dataset size, and other issues, and their predictive performances were never satisfactory [248–250].

In recent years, the advancement of computing power has driven the rapid development of AI, while positively promoting the development of computational chemistry and pharmacology. For example, various ML and DL methods were used in various Kaggle competitions to improve the predictive performance of QSAR methods, all of which achieved high performance [78]. As mentioned above, DL allows the identification of new molecular representations instead of relying solely on off-the-shelf and expert-derived chemical signatures. AI algorithms relying on rich biomedical data show promising prospects in areas such as bioactivity prediction, VS of drugs, and *de novo* drug design.

Before going into details, it is necessary to briefly introduce the concepts of structure–activity relationships (SARs) and QSARs. These two concepts are frequently used in drug design using ML and DL methods and are powerful aids in the design, optimization, and development of drugs. SARs are based on the assumption that molecules with similar structures have similar activity. In drug discovery, QSARs are based on various molecular characterization methods (e.g., molecular descriptors and molecular fingerprints) and mathematical models to describe the mathematical relationship between the structure of a molecule and its specific biological activity. A QSAR model assumes that the structure of a compound determines its physicochemical properties and biological activity; therefore, quantitative relationships can be established between the structure of a compound and its physicochemical properties, biological activity, toxicological effects, and so forth. The QSAR analysis process usually includes the preparation of preliminary datasets, the calculation and selection of molecular descriptors, the establishment of relevant models, and the evaluation and validation of model results [248,251].

4.1. Cutting-edge techniques facilitating VS

VS has endured for the past decade or so. In order to reduce the number of compounds that actually need to be measured and increase the efficiency of lead compound discovery, the *in silico* approach is used to simulate the interaction between a target and a small molecule and predict the affinity between the two before a bioactivity test is performed [252]. VS methods are often classified into structure-based VS (SBVS) or ligand-based VS (LBVS) [253–255]. The combination of AI and VS has brought a new dynamism to the field. A variety of molecular characterization approaches combined with various novel model architectures have provided new insights into the discovery of new compounds [9].

SBVS selects potential ligands based on the 3D conformation of the protein and scores the ligand's ability to bind to the protein based on the inputted knowledge of biophysical methods, resulting in a ranking of drug candidates. Previously, simulations using various docking software were the dominant approach and resulted in many algorithms, such as Monte Carlo (MC) algorithms [256] and molecular dynamics (MD) algorithms [252,257,258]. A primary limitation of the simulation results is the construction of the scoring function, which must take many factors into account along with their plausibility as parameters. AI takes these many factors as features of the data, implicitly learns the relationship between the features and the experimental results, extracts useful nonlinear mapping relationships from them, and gives a final score. A VS method known as ID-Score [120] selected nine classes of property descriptors (i.e., van der Waals interaction, hydrogen-bonding interaction, electrostatic interaction, π-system interaction, metal–ligand bonding interaction, desolvation effect, entropic loss effect, shape matching, and surface property matching) as features, used 2278 compounds as the training set, and used a support vector regression (SVR) algorithm to fit the binding affinity of small molecules to proteins. The results showed that ID-Score can correctly distinguish structurally similar ligands, demonstrating its use as a powerful tool for assessing structure-based drug–protein affinity.

In another study, a CNN was used to score protein ligands. Unlike traditional methods, CNNs are powerful enough to accept 3D representations of protein–ligand interactions as input. During the training of the model, the CNN learns the key features affecting binding from the 3D representation, which is used to determine the correct or incorrect binding pose and known binders and non-binders. Xie et al. [259] took a different perspective to improve the efficiency of lead compound discovery by combining an SVM classification model with a docking-based VS method. More specifically, they developed an SVM model to distinguish inhibitors of the target from non-inhibitors and performed a docking-based VS on this basis. This combination greatly improved the hit rate and enrichment factor of the VS. In contrast to the work by Xie et al. [259], Pereira et al. [260] developed DeepVS, which uses a DL approach to optimize docking-based VS. In this study, a directory of useful decoys (DUD) [261] was used as the benchmark dataset to evaluate the method. Dock [262] and Autodock Vina1.1.2 [263] were used as docking programs to generate protein–compound complexes. Then, essential processing of the protein–compound complexes was done and the results were fed into the CNN model as input. The CNN model extracted the key features from this essential data and used them to evaluate the score of the ligands. The results showed that the proposed DeepVS achieved advanced performance on VS.

In comparison with the SBVS approach, which is limited by the structural information of the target protein, LBVS can make full use of the known ligand bioactivity data and screen a large database of compounds to discover potential lead compounds. Therefore, AI-based VS tends to favor LBVS. The starting point of LBVS is the assumption that structurally similar compounds have similar biological activities; thus, the AI methods used in this field include both regression models for activity prediction and classification models based on compound similarity.

QSAR is widely used in LBVS because of its use of mathematical models to relate molecular structures to quantitative biological activities. NB, RF, and SVM are very popular algorithms in LBVS. AbdulHameed et al. [264] screened a database with nearly 2000 compounds using a QSAR-based model with an NB algorithm and using the physicochemical properties of the molecules as features. Finally, it was found that activators of pregnane X receptor (PXR) tend to be hydrophobic, while the *in vitro* and *in vivo* activities are often consistent. Profile-QSAR 2.0 was presented to predict the activity of compounds [265]. Compared with the earlier profile QSAR (pQSAR) 1.0 method, the pQSAR 2.0 method used the historical activity values of the compounds as variables. The optimized pQSAR used an RF model to predict the IC50 values, achieving the same accuracy as the medium-throughput four-concentration IC50 measurements. Chen and Visco [266] created a pipeline inte-

Table 4
Toolkits and libraries for analyzing small molecules and proteins.

Toolkits	Introduction	Main tools	Refs.
RDKit	RDKit is a cheminformatics toolkit that offers a variety of implementations for studies using cheminformatics, including algorithms for creating molecular fingerprints, searching for molecular structures, and building 2D to 3D structures	Generating molecular fingerprints, molecular structure searching, 2D to 3D structure construction, generation and manipulation of chemical formats	[192]
OpenBabel	OpenBabel is a toolkit that provides a number of features, including the creation of multiple molecular fingerprint types, format conversion between different chemical formats, structure and substructure searches based on graph isomorphism, and tools for organic chemistry	Generation of several types of molecular fingerprints, conversion between various chemical formats, structure and substructure searching based on graph isomorphism, organic chemistry tools	[193]
DayLight toolkit	DayLight toolkit provides the creation of SMILES strings, graph-based substructure search, analysis of 2D and 3D chemical structures, and creation of several kinds of fingerprint	Generation of SMILES strings, graph-based substructure search, analyzing 2D and 3D structures of compounds, generation of different types of fingerprints	[194]
CDK	The CDK is a library that provides tools for creating and modifying different chemical formats, doing substructure searches, using graph theory techniques to find chemical structures, and creating 3D structures	Generation and manipulation of chemical formats, substructure searching, implementation of graph theory algorithms for chemical structure searching, 3D structure generation	[195]
OpenEye toolkit	OpenEye toolkit offers a variety of features, including handling representations of different chemical formats, shape similarity and grouping algorithms based on 2D structure, and 2D visualization of molecular structure	Handling representations of several chemical formats, 2D structure-based shape similarity, clustering methods and 2D molecular structure rendering	[196]
ChemmineR	ChemmineR is a framework for the R language and an environment for statistical computing, including implementations of algorithms for handling multiple compound representations, 2D structural similarity searching, compound library clustering algorithms, classification algorithms, and visualization techniques	Implementations of algorithms for handling different types of compound representations, 2D structural similarity searching, clustering algorithms for compound libraries, classification algorithms, visualization methods	[197]
Indigo	Indigo is a cheminformatics toolkit that offers a variety of utilities, including the ability to work with SMILES strings, use the SMARTS to search for structure and substructure, and create several kinds of fingerprints	Manipulation of SMILES strings, structure and substructure searching generation of various types of fingerprints	[198]
SHAFTS and eSHAFTS	SHAFTS is a fast 3D similarity calculation tool that adopts a hybrid similarity metric of molecular shape and label chemistry groups by pharmacophore features to calculate and rank the 3D similarity of small molecules. eSHAFTS provides a user-friendly graphical working environment based on SHAFTS	Fast 3D structural similarity searching for small molecules, provides a graphical user interface	[199–201]
PROFEAT	PROFEAT is a web server that uses the structural and physicochemical properties of amino acids to compute protein descriptors from input protein sequences. Biological networks, ligand–protein interactions, PPIs with ligands, and ligand–protein interactions are all computed as descriptors as well	Computing protein descriptors	[202]
BLAST	BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance	Pairwise sequence alignments/database search	[203]
ClustalW	ClustalW is a program for multiple sequence alignment that uses guide trees to compare sequence similarities between three or more sequences. It is a very sensitive progressive technique that can be applied to divergent protein sequences and makes use of sequence weighting and position-specific gap penalties	Multiple sequence alignments	[204]
Dali	Dali is an online server that evaluates protein 3D structures based on distance matrices. It accomplishes comparisons against PDB structures as well as pairwise and all-against-all multiple comparisons. It calculates RMSD scores but assesses similarities using Dali *Z*-scores, where scores greater than 2 indicate a substantial structural similarity	Protein structure alignments	[205]
MultiProt	MultiProt is an online structural alignment tool that can be used for the simultaneous alignment of multiple protein structures. It does not require the alignment of all the input molecules; instead, it finds the common geometrical cores between them by detecting high-scoring partial multiple alignments	Simultaneous alignment of multiple protein structures	[206]

(*continued on next page*)

Table 4 (*continued*)

Toolkits	Introduction	Main tools	Refs.
TM-align	TM-align is a structural alignment tool that is used to compare protein structures without using sequence order information. It combines dynamic programming iterations with a TM-score rotation matrix	Comparison of protein structures	[207]
PocketShape	PocketShape is a 3D structure-based evaluation method for exploring protein binding site similarity; it depends on structure-based alignment and is capable of detecting common patterns or residue reservation between binding sites, which are not required to possess continuous residues or be homologous	3D similarity calculation of protein binding sites	[208]
RCSB PDB	RCSB PDB comparison tool is an online tool that computes pairwise sequence comparisons using the blast2seq, Needleman–Wunsch or Smith–Waterman algorithms and pairwise structure comparisons using the FATCAT, CE, Mammoth, TM-Align, or TopMatch algorithms	Pairwise structure/sequence alignments	[209]
SiteEngine	SiteEngine is a structural similarity measurement tool created for the prediction of potential protein binding sites based on the similarity of their geometrical and physicochemical properties with known binding sites	Prediction of potential binding sites of proteins	[210]
APoc	APoc is a large-scale, sequence-order-independent structure alignment tool for comparing experimentally confirmed or computationally predicted protein pockets. It makes use of an algorithm that creates initial alignments from gapless alignments, secondary structure alignments, fragment alignments, and local contact pattern alignments	Comparison of predicted protein pockets	[211]
eMatchSite	eMatchSite is a sequence-order-independent local alignment tool that is used to align and compare the ligand binding sites of computationally generated protein models. It makes extensive use of evolutionary data gleaned from entropy and secondary structure profiles of weakly homologous templates in complexes with ligands	Ligand binding sites	[212]
FragHMMent	FragHMMent is a bioinformatics tool that predicts residue–residue contacts in a protein sequence. It makes use of local protein structure descriptors, predicted secondary structure, and HMMs trained on homologous sequences	Prediction of residue-residue contacts	[213]
PSIPRED	PSIPRED is a web server that predicts secondary structures of proteins from their amino acid sequences	Prediction of protein secondary structure	[214]
SCREEN	SCREEN is an online tool that locates and describes protein surface cavities. It creates a set of cavities for each structure and calculates each cavity's geometric and electrostatic properties	Predicts cavities of protein surfaces	[215]

CDK: chemistry development kit; SMARTS: simple modular architecture research tool; RMSD: root-mean-square deviation; HMMs: hidden Markov models;

Table 5
Popular methods for DTI identification and the algorithms applied.

Method	Abbreviations	Algorithms	Description	Ref.
Similarity/distance-based methods	SITAR	Similarity-based inference of drug targets	A new scoring method for drug–gene associations based on drug–drug and gene–gene similarity measures	[226]
	MDTI	MultiviewDTI	A model based on normalized spectral clustering that integrates drug similarity network data and target similarity network data from two views	[230]
	LPLNI, LPLNI-II	Label propagation method with linear neighborhood information	An algorithm for predicting novel interactions between known drugs and targets by using drug–drug linear neighborhood similarity and old DTIs	[231]
	WNN-GIP	Weighted nearest neighbors-Gaussian interaction profile (GIP)	An algorithm that uses a weighted nearest-neighbor procedure to infer the profile of a new drug from the interaction profile of a known compound	[232]
Feature-based	BGL	Bipartite graph learning	A supervised learning model of a bipartite graph that integrates chemical and genomic spaces into a unified space called the "pharmacological space"	[221]
	Boosting	SimBoost, SimBoostQuant	A method that can predict continuous values of drug and target binding affinity, thereby integrating the overall interaction profile from true-negative to true-positive interactions; SimBoostQuant evaluates the confidence of predicted affinity by computing prediction intervals to determine the domain of applicability metrics	[233]
	BE-DTI	Bagging-based ensemble method	A DTI prediction model using dimensionality reduction and active learning	[234]
Matrix factorization	NRLMF	Neighborhood regularized Logistic matrix factorization	A DTI prediction method combining logical matrix factorization and neighborhood regularization	[235]
	PMF	Probabilistic matrix factorization	A DTI prediction model that uses a collaborative filtering algorithm and does not require 3D shape similarity	[227]
	VB-MK-LMF	Variational Bayesian multiple kernel logistic matrix factorization	A DTI prediction model that simultaneously possesses the advantages of multi-kernel learning, weighted observations, Laplacian regularization, and explicit modeling of binary DTI probabilities	[236]
	LRE, SLRE, MLRE	Low-rank embedding	A method for inpainting and minimizing reconstruction error in the embedding space via the low-rank representation of the dataset, thereby preserving point-wise linear reconstruction; SLRE is an arbitrary-view-based low-rank embedding single-view method, and MLRE is a multi-view-based method	[237]
Network-based methods	NBI	Network-based inference	A DTI prediction model using topological similarity of drug–target bipartite networks	[238]
	NRWRH	Network-based random walk with restart on the heterogeneous network	A DTI prediction method that aggregates three different networks into heterogeneous networks and implements a random walk	[239]
	NetCBP	Network-consistency-based prediction method	A semi-supervised learning model utilizing both labeled and unlabeled interaction data	[240]
	DTINet		A DTI prediction computational pipeline capable of parsing topological properties of drug nodes in heterogeneous networks	[223]
Hybrid methods	DT-Hybrid	Domain-tuned hybrid	An NBI approach with domain-based knowledge including drug and target similarity	[241]
	MGRNNM	Multi-graph regularized nuclear norm minimization	A DTI prediction method incorporating multiple graph Laplacians on drugs and proteins into the framework	[242]
	RBM	Restricted Boltzmann machine	A two-layer graphical model that integrates multiple types of DTI and predicts unknown DTI	[243]
	LRF-DTI	LASSO-based RF method	A DTI prediction method that inputs the lasso-processed feature vector into an RF classifier	[244]
DL	DeepDTIs	DL in predicting DTIs	A DL approach to predict the interaction of a new drug with an existing target or a new target with an existing drug using unsupervised pre-training to extract representations from raw input descriptors	[245]
	DeepConv-DTI	DL with convolution-DTI	A DL approach for large-scale prediction of DTI using CNN on raw protein sequences to capture local residue patterns of proteins	[246]
	DeepDTA	Deep DT binding affinity prediction	A DL-based model that predicts DTI binding affinity using only sequence information for proteins and drugs	[193]
	DeepTrans	Deep transcriptome data	A DL model modeled as a binary classification task that predicts potential DTIs using transcriptome data from the Lincs-L1000 database	[247]

grating QSAR with an SVM model to identify the inhibitors of Cathepsin L. They used a signature—a descriptor based on fragments—as the model's input. After optimizing the model, nine out of 12 screened compounds were experimentally confirmed. ANNs are another commonly used tool in QSAR studies. Myint et al. [267] reported an ANN-based QSAR method called fingerprint-based ANN (FANN)-QSAR that uses three different molecular fingerprints: ECFP6, FP2, and MACCS. The well-trained model was used to predict the affinity of cannabis ligands and found compounds with a good affinity for CB2. In another group study, the minimal inhibitory concentration (MIC) of quinolones was determined by using topological descriptors in an ANN [268]. As more DL methods have gradually been used for QSAR-related studies, researchers have found that DL tends to outperform ML in both single-task and multi-task learning [269–271].

QSAR methods are not the only tools used for LBVS [272–274]. Li et al. [275] used multiple ML methods to construct classification models to select liver X receptor (LXR) agonists. During this process, optimized property descriptors and topological fingerprints were used to characterize small molecules in the database and constitute a total of 324 models with four algorithms: NB, SVM, KNN, and recursive partitioning (RP). The top 15 models were selected for evaluation, and ten models were found to have an accuracy of more than 90%. In another study, an SVM with NB was used to identify butyrylcholinesterase (BuChE) inhibitors [276]. Initially, 1870 descriptors were selected; after analysis, activity-related descriptors were then selected to reduce noise. A better performance was eventually achieved. There are also numerous examples of self-organizing mapping (SOM) being used in LBVS [277]. For example, Hristozov et al. [278] used SOM as a model to recognize and exclude compounds that are unlikely to have specific biological activity. The power of SOM has also led to its use in some software [279].

With the rapid increase in the number of known compounds in recent years, DL architecture has been found to be more suitable for processing large compound datasets. One group trained with existing HTS data and used a molecular graph as input to a neural network to learn molecular representations [280]. Compounds with similar representations were then assigned in the neighboring hyperdimensional feature space. After learning the features, the similarity to drug molecules in a large compound library was measured using cosine similarity, and the small molecules in the library were ranked and filtered to obtain lead compounds. Unlike the use of graph models to generate the features of small molecules, adversarial AEs (AAE) were used by Kadurin et al. [281] to construct a small molecule feature generator. Based on the obtained features, 72 million compounds in PubChem were screened to discover potential anticancer drug molecules. CNNs are widely used in image recognition; thus, for the purpose of using CNN models in drug research, molecules or proteins are often characterized in the form of matrices. Xu et al. [282] directly used images of molecules as input to CNN models to screen for inhibitors of chemistry development kit 4 (CDK4) and achieved better effects than competing models. The use of DL for LBVS has been increasingly studied in recent years, and models such as RNN [283] and RL [284] have been used for drug discovery, providing more opportunities and benefits for LBVS.

Overall, efficient lead compound discovery through VS is still a huge challenge, as there is no satisfactory way to address issues such as the activity cliff. AI algorithms are powerful tools that can be used not only for SBVS but also for LBVS to help break through the relevant challenges and assist in *de novo* drug design. As the complexity of algorithms increases and high-quality data becomes available in future, bottlenecks in existing technologies will continue to be broken, facilitating the discovery of new drugs.

4.2. Recent progress in de novo drug design

The aim of drug design is to design drugs with specific properties that satisfy specific criteria, including efficacy, safety, reasonable chemical and biological properties, and structural novelty. In recent years, *de novo* drug design with the help of deep generative models and reinforcement learning algorithms has been considered to be an effective means of drug discovery. This approach can bypass the drawbacks of the traditional empirical-based drug design paradigm and allow computers to learn the drug targets and molecular features by themselves to generate compounds that meet specific requirements at a faster and less costly rate [285–287].

De novo drug design according to protein structure used to be the dominant approach. In this approach, whether designing new molecules directly from protein structures or making reasonable inferences from the properties of known ligands, the corresponding ligands are designed according to the spatial and electric potential constraints of the target protein binding pocket in order to discover molecules with specific properties. A huge limitation of these early approaches was that the resulting new molecules were not chemically accessible—that is, their structures were practically impossible to synthesize or extremely difficult to produce, or the molecules had poor druggability. In addition, many *de novo* drug design approaches utilize fragments of molecules with known properties for molecular assembly, and use large libraries of molecular fragments to generate and design molecules with novel structures while ensuring that the molecules can be synthesized. However, this approach relies on chemical knowledge to replace or add molecular fragments, which will restrict the search space and ignore certain potential molecular structures. The generation of new molecules with deep generative models and the targeted optimization of models with reinforcement learning algorithms can solve the problems of the above traditional methods in a more satisfactory way [288–290].

Deep generative models are of great advantage in the field of *de novo* drug design, as they do not require explicit prior input of chemical knowledge during the generation of molecules. These models can search in a broader unknown chemical space to automatically design novel molecular scaffolds beyond the limitations of existing molecular scaffolds. Deep generative models that are widely used for *de novo* drug design include RNN-based generative models, variational AEs, AAEs, and GANs. The process of designing molecules with generative models is highly stochastic, and the generated molecules are highly variable in structure and uneven in quality. Reinforcement learning can enable generative models to perform targeted optimization by fine-tuning the model parameters so that the generated molecules have specific drug molecule properties.

RNN-based generative models can generate compounds with similar biochemical properties as the sample compound but with a completely new scaffold structure. The training process starts by using a large chemical database to train the RNN model so that the model can learn how to generate the correct chemical structure. Reinforcement learning algorithms are then used to fine-tune the RNN parameters so the model is capable of mapping generated chemical structures to a specified chemical space. Reinforcement learning enables the RNN-based generative model to generate new molecules with promising pharmacological properties, while ensuring the structural diversity of the generated molecules. A single reinforcement learning reward mechanism often leads to relatively simple structures of the generated molecules, so an appropriate and multi-perspective reward function must be selected to guide molecule generation. Olivecrona et al. [123] developed a sequence-based approach to *de novo* drug design

called REINVENT. First, the researchers collected 1.5 million molecules from the ChEMBL database that satisfied specific requirements and used SMILES of these molecules to train the RNN model to learn the characteristics of active molecules and generate new molecules. The generated molecules were then scored using a reinforcement learning algorithm to fine-tune the RNN parameters, so that new compounds with activity against a specific target could be generated. This method was applied to several different molecule generation tasks in the study, including the generation of sulfur-free molecules, backbone expansion from a single molecule to generate celecoxib-like structures, and the generation of new inhibitor molecules for type 2 dopamine receptors.

Another area in which RNN-based generative models are applied in drug design is the optimization problem of lead compounds [291]. A new molecular generation algorithm called scaffold-constrained molecular generation (SAMOA) was proposed to solve the scaffold constraint problem within the lead compound optimization problem. The study used an RNN generation model to generate SMILES sequences of new molecules, and then used a refined sampling procedure to implement the scaffold constraint and generate molecules. A strategy-based reinforcement learning algorithm was also applied to explore the relevant chemical space and generate new molecules matching the expected ones. The DeepFMPO framework proposed by Ståhl et al. [292] started from an initial set of lead compounds and modified the structure of these lead molecules by replacing some of their fragments. This study confirmed the wide use of RNN-based generative models in the field of molecular generation.

As deep generative models, VAEs are often used in various generative tasks, including the *de novo* design of small molecules and the generation of peptide sequences. A group constructed a molecular generation model based on a conditional VAE for *de novo* molecular design with a three-layer RNN for both the encoder and decoder. The results demonstrated that this model can design drug-like molecules with five target properties and can also tune individual molecular properties without affecting other properties [124].

In 2019, Insilico Medicine published a study [28] on the rapid *de novo* design of potent discoidin domain receptor 1 (DDR1) kinase inhibitors using a VAE. Several new compounds with inhibitory activity against DDR1 kinase were identified, chemically synthesized, and experimentally validated in just 21 days. This study demonstrated the potential of the method to perform fast and efficient molecular design. The generative tensorial reinforcement learning (GENTRL) model consists of two main components: a VAE and a strategic gradient reinforcement learning algorithm. The VAE is used to generate new molecules, while the reinforcement learning fine-tunes the model parameters to make the new molecules generated by the VAE more consistent with the expected properties. The encoder of the VAE is used to encode known molecules into hidden vectors. The decoder samples and decodes the hidden vector into a new molecule based on the hidden vector space. A reinforcement learning algorithm is used to guide the VAE-directed optimization during the training process. After model construction, Insilico Medicine used GENTRL to generate four new active compounds, two of which were validated in cellular experiments. Moreover, one of the lead compounds was tested in mice and was shown to have good pharmacokinetic properties. This study provides strong evidence that reinforcement learning combined with deep generative models can accelerate the process of and provide new insights into *de novo* drug design.

GANs are capable of generating new samples with a similar distribution to real data and have advantages in the fields of image recognition and natural language processing (NLP). In the pharmaceutical field, GANs are often integrated with techniques such as feature learning and reinforcement learning, and have played an important part in protein function prediction, small molecule generation, and more. Various molecular generation models have been constructed based on GANs, such as Mol-CycleGAN [293], objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC) [294], and reinforced adversarial neural computer (RANC) [295]. ORGANIC is a well-known molecular generation model that has become a comparative baseline model for other models. Its combination of a GAN model and a reinforcement learning algorithm can generate novel and effective molecules. The molecule generation performance of the RANC model has surpassed ORGANIC in many aspects, including the ability to generate new molecular structures and drug-like properties of molecules, which allows the design of active new molecules for different biological targets and covers a wide chemical space.

In addition, Harel and Radinsky [296] proposed a molecular template-driven neural network that combines a VAE, CNN, and RNN to generate chemical structures with similar properties to the template molecules while being structurally diverse. The researchers found that the proportion of effective molecules among the generated molecules was significantly enhanced by adjusting the sampling process of the VAE.

Molecules designed by computer must not only have good physicochemical properties but also be highly active and selective for the target under study; therefore, the question of how to set up an effective reward function is an important challenge in reinforcement learning. A combination of the framework of deep generative models with reinforcement learning algorithms drives the development of the drug design field and will have significant applications in the future in the *de novo* design of small-molecule and peptide drugs.

4.3. Application of advanced techniques in antibody design

Due to the wide application of ML and DL in chemistry, biology, and medicine, as well as their use in basic research in various fields, researchers now have a profound comprehension of biomolecules and systems biology. In the future, the direction of drug R&D will be biased toward the research of small molecules; moreover, bio-innovative drugs will gain ground. Similarly, there are already many DL approaches for the study of biological macromolecules drugs, both now and in the near future, such as oligonucleotides, monoclonal antibodies, or peptides with specific pharmacological properties. Here, we will elaborate on the design of antibodies.

Since antibodies are inherently biological macromolecules, the characterization of antibodies is similar to the encoding of proteins and RNAs. There are six general strategies for encoding antibodies: "one-hot" encoding, substitution matrix, amino acid properties, learned amino acid properties, encoding of supplementary attributes, and encoding of structural features [297]. The application of AI in antibodies is different from its application in ordinary biomolecules because antibodies are biological agents that can be used for disease treatment. Therefore, the design of antibodies has more in common with the design of drugs, since safety and efficacy of drugs must be taken into account. At present, AI-based methods are often used for antibody structure prediction, antigen–antibody binding prediction, antibody generation/design, deimmunization studies, and antibody sequence-based studies [297].

The AlphaFold2 DL system has been able to solve most of the protein structure prediction problems; however, for antibody structure prediction, as a special subfield of protein structure prediction, it is necessary to capture the subtle differences in the structure with extreme precision. Many methods have been developed to solve this problem, such as DeepAb [298] and DeepH3 [299]. To perform VS for the binding of antibodies to target antigens, a structure-based framework called DL for antibodies (DLAB) was

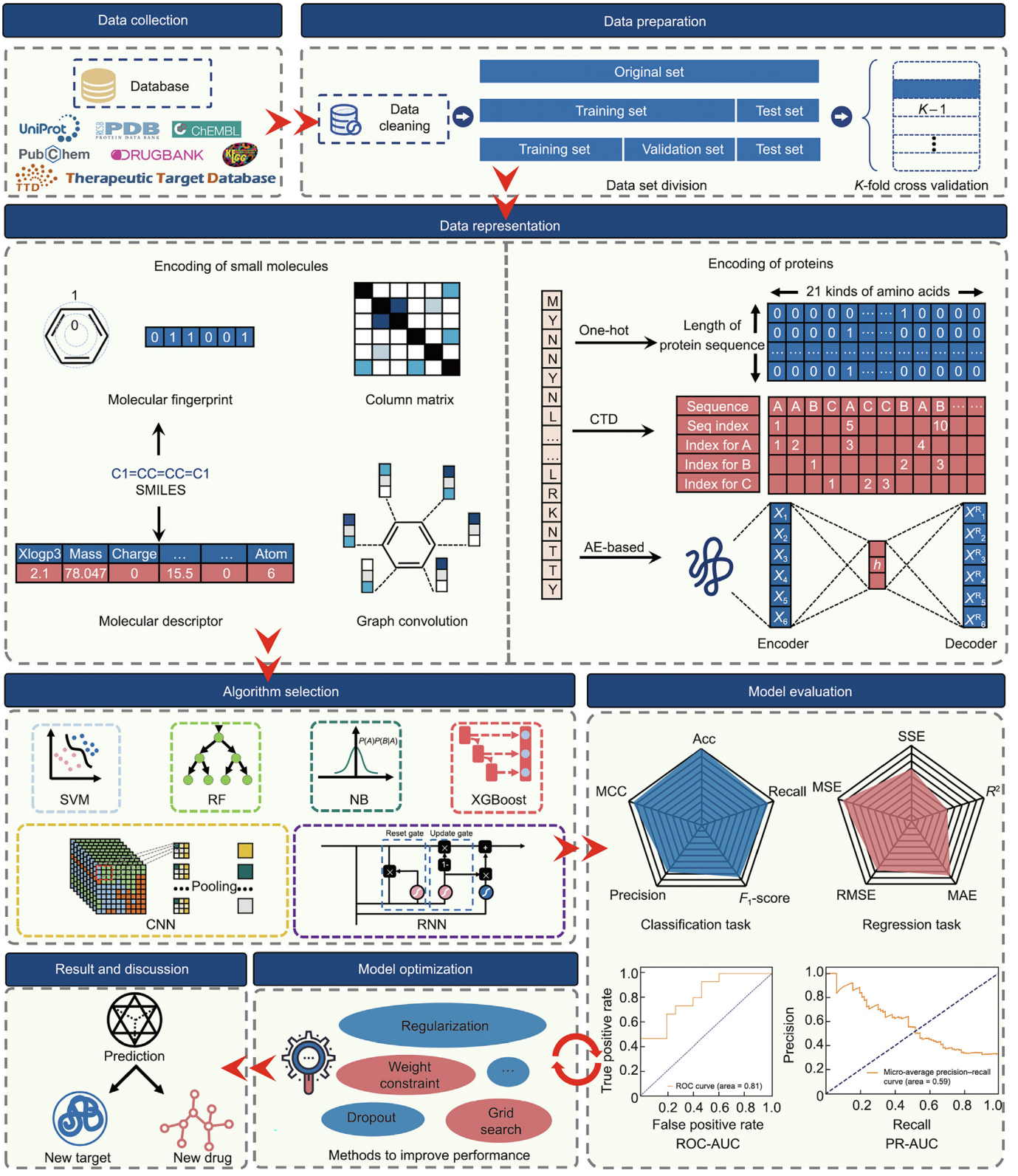

Fig. 3. A brief workflow of new target and drug discovery based on AI. First, useful data required for modeling and evaluating the model is collected and initially processed; it is then divided into a training set, validation set, and test set. Next, data in different formats is encoded as vectors or matrices for input to the model. The prepared data can be represented in various ways (e.g., small molecules can be represented by molecular descriptors, molecular fingerprints, and graph-based representations, while proteins can be represented by sequence correlation features (PSSM, AAC, CTD, etc.), AE, pre-trained protein language models, etc.). Depending on the problem to be studied, an appropriate algorithm must be selected to perform the prediction task. Criteria are adopted to evaluate the performance of the proposed model; according to these criteria, it is necessary to continuously adjust the parameters of the model and apply tricks to improve the performance of the model. Finally, a reasonable discussion and analysis of the prediction results are required.

proposed to improve antibody–antigen dockings [300]. As DLAB is a structure-based approach, it can optimize the pose ranking of antibody docking experiments and select antibody–antigen pairs for which accurate poses are generated and properly ranked. This approach has also demonstrated that the SBVS of antibodies can strongly complement traditional experimental screening methods.

The search for new antibody sequences is a major research hotspot in antibody discovery. Early computational approaches attempted to use enumeration methods for new sequence discovery and subsequent prediction work. Although these methods reflect the diversity of designed antibodies, they do not explain these discoveries in a biological sense and lack conviction. Recently, the potential features of antibodies—including the frequency of amino acid positions and the physicochemical properties of the antibody—have been learned by GANs or VAEs [301]. These methods provide a new way of thinking and a new approach for antibody generation and design, which can be relied upon in the future to design therapeutic antibodies via DL.

The directions for the development of antibody drugs discussed above stem from a starting point that is similar to that of the design of small molecule drugs. Antibodies can be designed differently than traditional drugs due to their large molecular weight and attributes such as biomolecular function. In designing an antibody drug, it is necessary to consider the immune response the drug elicits when it enters the body. Thus, it is critical to use ML algorithms for analysis of next-generation sequencing (NGS) data to carry out deimmunization studies of antibodies [302]. In addition, antibodies similar to human antibodies must be designed without loss of activity during the humanization process. [303]. Novel humanization (e.g., Sapiens) and humanness evaluation methods (e.g., OASis) are two data-driven approaches to address these issues. Sapiens uses a masked language modeling (MLM) model to learn the humanization method of antibodies, while OASis is used to evaluate the humanness of an antibody sequence. BioPhi successfully combined these two algorithms to capture the intrinsic features of antibody complexes and provide similar mutation selection to that used experimentally for humanized mutations. This achievement indicates that DL will be indispensable in the deimmunization studies of antibodies. Another major feature of DL in antibody research is its ability to use NLP to learn and encode the antibody space to reveal new insights into the biological function of antibodies. For example, antibody-specific bidirectional encoder representation from transformers (AntiBERTa) [304] and AbLang [305] can understand the back-and-forth association of antibody sequences and, based on this understanding, can infer specific masked regions.

When conducting antibody drug research, DL can be used to connect the microscopic properties of molecules with the macroscopic results of experiments and provide additional insights into the biology associated with immunoglobulins. Therefore, DL approaches are increasingly being applied in the research and design of therapeutic antibodies to enable the efficient development of new antibodies and provide a new strategy for the future pipeline of antibody design. Overall, AI has shown promising power in drug target identification and new drug discovery. Fig. 3 depicts a generic workflow using AI for target and drug identification.

5. Application of AI to preclinical drug research

Preclinical studies focus on non-clinical pharmacology, pharmacokinetics, and toxicology studies. The physicochemical properties of a drug and its ADMET properties are essential for pharmacokinetic and toxicology studies [33,306]. Unsuitable properties of drug candidates will lead to the failure of the expensive drug development phase [307]. The failure rate and loss of clinical studies can be decreased by early evaluation of the relevant properties of drug candidates.

5.1. Prediction of physicochemical properties

The ADMET properties of a drug candidate can be directly influenced by its physicochemical properties and will have a critical impact on the success of a drug entering the market [308,309]. For example, the ionization constant (pK_a), which is the fundamental parameter underlying properties such as logD and solubility, affects the aqueous solubility of a molecule, which can in turn affect the drug formulation method. Moreover, the ADMET of compounds under different pH conditions are profoundly influenced by the charge state of the compounds [310]. Although lead compounds with promising drug-like properties may not always be successfully marketed, promising properties are still an inspiration for drug design. However, physicochemical properties are not easily measured directly, and accurate prediction of the properties of small molecule drug candidates facilitates further structural optimization of small molecules until they are designed to meet the desired properties.

Some approaches for predicting the physicochemical properties of molecules focus on predicting a certain physicochemical property, such as lipophilicity [311] or aqueous solubility [312], while others predict several physicochemical properties together [99]. Although molecules can be represented in a variety of ways, predictions for a single property may use certain specific features, such as the number of hydrogen bonds [313] and the connectivity indices of various molecules [314] correlated with solubility. To date, accurate prediction of the aqueous solubility of small molecules remains a challenge [315], but DL methods have been found to be more effective than previous ML methods in this endeavor [316]. In the Second Challenge to Predict Aqueous Solubility, one of the models [317] combined an NLP approach to obtain embedding vectors based on small molecule SMILES, in order to feed these vectors into the transformer model for predicting molecular aqueous solubility. Francoeur and Koes [317] found that overly complex models did not perform as well as small DL models in this task, which may be due to overfitting of the model as a result of the complex model and the smaller amount of data.

To address the issue of simultaneously predicting several physicochemical properties of small molecules, researchers have focused on molecular feature learning and characterization; examples include molecular feature learning and representation based on a GNN architecture [98], combining traditional molecular representation approaches with features learned by message-passing neural networks (MPNNs) [99], and a form of graphical representation of molecular design based on extended-connectivity circular fingerprints (ECFPs) [318]. Shen et al. [319] proposed a new form of molecular representation that involved first calculating the distance matrices of molecular fingerprints and the molecular descriptors of eight million molecules, respectively, and then reducing the distance matrices to two dimensions via uniform manifold approximation and projection (UMAP) to form a scatter plot. Next, the dimensionality-reduced scatterplots were assigned to 2D grid maps using the Jonker–Volgenant (J–V) algorithm. Finally, the data was divided into different channels based on different molecular fingerprints or descriptors. These molecular representation forms were fed into a CNN for the prediction of molecular properties, achieving a SOTA performance on multiple datasets.

5.2. Prediction of ADMET-related properties

The failure of most clinical trials is often blamed on inadequate ADMET studies of the drug, rather than on a lack of certain efficacy.

The "absorption, distribution, metabolism, excretion (ADME)" portion of ADMET often determines whether a drug molecule will reach the target protein *in vivo*, what protein will transport or metabolize this drug [47,320], how long it will stay in the blood, and when it will be inactivated, while the "T" portion (i.e., toxicity assessment) is a major concern in phase I clinical trials. If the risk of clinical trial failure can be reduced via thorough preliminary ADMET studies, significant money and time costs will be avoided [321,322]. With hundreds of compounds waiting to be evaluated for their ADMET properties in the early drug discovery phase, it would be time-consuming and expensive to validate each one through extensive animal studies. Therefore, the use of AI to rapidly and accurately predict the ADMET properties of drugs has been widely adopted [323].

QSAR and quantitative structure–property relationship (QSPR) models play pivotal roles in the ADMET prediction of small molecules. Many ML methods, in combination with QSAR or QSPR models, have performed well in ADMET prediction [324]. Most of these ML methods focus on several ADMET properties [325], such as human ether-a-go-go related gene (hERG)-mediated cardiotoxicity [326], blood–brain barrier penetration [327], permeability glycoprotein (P-gp) [328], cytochrome P450 (CYP) enzyme family [329], acute oral toxicity [330], carcinogenicity [331], mutagenicity [332], respiratory toxicity [333], or irritation/corrosion [333]. Zhu et al. [334] used a QSPR model to predict the blood–brain partition coefficient (logBB). The researchers used four ML methods—namely, SVM, multivariate linear regression, multivariate adaptive regression splines, and RF—to predict this property for 287 compounds and found that the polar surface area and octanol–water partition coefficient were strongly relevant to the blood–brain partitioning. A CYP enzymes-inhibition prediction model based on the C5.0 algorithm (a decision tree model algorithm) was constructed using several molecular fingerprints or molecular descriptors as inputs to predict five CYP enzymes related to drug oxidation or hydrolysis [335].

Most of the ADMET datasets are imbalanced and have high dimensionality problems, and the integrated learning approach has been applied to deal with these two types of problems. The processing of imbalanced data, the combination of multiple models, and optimization steps have been integrated to form an adaptive ensemble classification framework (AECF) [336]. Yang et al. [336] used AECF to predict a variety of ADMEproperties using multiple ML methods; their results all had satisfactory AUROC values ranging from 0.78–0.91. This ensemble approach was demonstrated to be a very useful multi-classification system through validation with the DrugBank database.

DL approaches are also widely applied to the prediction of ADMET properties. For example, a classical feed-forward backpropagation neural network (BPNN) architecture and a repeated double cross-validation (rdCV) approach were combined to estimate the blood–brain barrier penetration [337]. DL allows a model to be trained using a larger and more representative dataset, ensuring that a wider variety of compounds are covered than is possible with ML. Validated with external datasets, this method predicts values that are in good agreement with many experimentally derived logBB values. In another work, it similarly demonstrated that neural networks generally outperform ML methods for ADMET properties prediction. Montanari et al. [121] predicted seven different ADMET properties corresponding to each of the following endpoints: logD, solubility, melting point, membrane affinity, and human serum albumin binding. Moreover, Wang et al. [338] developed a DL model to predict drug metabolites with an accuracy superior to the popular rule-based method systematic generation of potential metabolites (SyGMa). In a comparison of a multi-task graph convolutional model, a fully connected neural network, and an RF model, it was shown that the multi-task graph convolutional model performed the best. However, for more complex tasks, such as the prediction of Caco2 permeation or *in vitro* metabolic stability, multi-task graph convolutional networks were unable to achieve good results, probably due to the simplicity of the model constructed in this study, which hindered the model from learning the deeper features. In addition, the multitasking model in this study was considered a trial-and-error exercise, and there were no specific experiences and rules about which tasks should be combined together.

Other recent work has similarly demonstrated the potential of multitasking models for ADMET properties prediction. Various user-friendly ADMET software and web servers have been developed for predicting the ADMET properties of molecules [125,339–342]; among these, ADMETlab 2.0 [125] is widely praised. ADMETlab 2.0 is based on a multi-task graph attention (MGA) framework and can predict multiple ADMET properties of drugs (it contains a total of 88 relevant parameters with 23 ADME properties, 27 toxicity endpoints, and eight toxicophore rules). Most of the data used for training was derived from bioactivity data in the open-access database, relevant literature, and toxicity prediction software (toxicity estimation software tools (TEST)). Based on these training sets and the novel model architecture, some of the properties predicted by ADMETlab 2.0 are unique in comparison with the results of similar tools. It is a convenient tool for non-expert users while being able to provide comprehensive and accurate ADMET properties for target molecules for medicinal chemists.

6. AI-assisted clinical trial design, post-market surveillance, and prognosis prediction

A drug candidate can be sent to clinical studies only after it has undergone the process from target identification to drug design, synthesis, and optimization, and then to preclinical studies of ADMET-related properties, which initially confirm the safety and efficacy of this compound. The clinical trial phase consumes most of the time and investment during drug R&D. Although AI cannot be used to directly predict the clinical trial results of drug candidates in clinical studies, it can be used to assist in the design of clinical trials to enhance the rationality and safety and ultimately provide a more realistic response to the clinical trial results of a drug. After phase III clinical trials, drugs also require long-term regulatory work to further identify undocumented toxic effects in previous studies in order to prevent malignant events.

6.1. AI-assisted clinical trial design

The high failure rate of clinical trials makes this the most difficult step in the new drug development pipeline, with about 90% of drug candidates being eliminated in clinical trials [343], where each failed clinical trial costs approximately 0.8 to 1.4 billion USD. To overcome these shortcomings, a number of AI-based approaches are now available to assist in crucial steps in clinical trial design, such as helping to improve patient recruitment and enhance patient monitoring [344]. To address the issue of patient selection, AI can be used to explore the association of patient biomarkers with external indications to predict the likely treatment response of patients, which can help in screening for patients with high clinical success [345]. In addition, e-phenotyping can be used to reduce patient population heterogeneity [346] and to aid patient selection through prognostic or predictive enrichment [347,348].

Patient monitoring in clinical trials is also a critical process. By incorporating wearable technology, AI can be used to help automate and personalize real-time patient monitoring, thereby reduc-

ing patient workload and improving medication adherence issues. Accurate medication adherence data can better reflect the results of clinical trials, and AiCure [349]—a new AI platform used to measure medication adherence—has shown a 25% improvement in adherence compared with traditional therapies in a phase II trial for schizophrenia. In addition, AI has been used to optimize dosing to reduce adverse effects, improve the safety of trial protocols, and reduce patient defaults due to safety concerns [350].

6.2. AI-assisted post-market surveillance and prognosis prediction

After a drug is approved and successfully enters the market after the clinical phase, it undergoes a long-term investigation to further monitor and evaluate the drug safety. Electronic health record (EHR) mining is an important data source for AI applications in post-market surveillance, in which the use of structured data can simplify the process of data pre-processing. Existing methods used in EHR include the self-control case series (SCCS) model [351], cohort and case-control methods [352], and temporal pattern-discovery algorithms [353].

Convolutional SCCS (ConvSCCS) is a scalable model for predicting longitudinal features using SCCS. Morel et al. [354] used step functions and exposures to avoid the problem of classical SCCS models that require a precisely defined risk window. The results showed a significant improvement in the computational speed and accuracy of the method and enabled its application to adverse drug reactions (ADRs) detection in a cohort of diabetic patients. Aside from the application of structured data, unstructured data from biomedical and clinical corpora can be used for NLP methods for DDI detection and classification [355] and the prediction of drug ADR [356]. Systems pharmacology, which is based on systems biology, studies the effect of drugs on the system as a whole; it is a rich source of data and is a common approach for AI in ADR mining. Lorberbaum et al. [357] proposed a network-based algorithm involving the modular assembly of drug safety subnets (MADSS). They combined systems pharmacology models with pharmacovigilance (PV) statistics to validate the algorithm, and the results showed a significant improvement in the prediction of adverse effects for four drugs.

Disease prognosis is the prediction of the course and outcome of the future development of a disease. In the past, clinicians usually relied on professional experience and traditional statistical analysis for clinical prognosis prediction, making it difficult to provide accurate results. Now, through the introduction of AI technology, multi-patient and multi-factor data can be analyzed to improve the accuracy of prediction results. In cancer prognosis, patient survival and disease recurrence are usually predicted. Enshaei et al. [358] used an AI model to compare the prediction accuracy of an ANN with traditional statistical methods (e.g., LR); the results showed that AI has higher accuracy in predicting the prognosis of OvCa patients. Nowadays, there are many ML and DL methods for the prognosis of various cancers, such as BrCa [359–363], lung cancer [364,365], gastric cancer [366–368], bladder cancer [369,370], and prostate cancer [371,372], illustrating the potential of AI technology in cancer prognosis.

7. Automation of drug synthesis with AI

The development of a new drug usually involves four stages: design, make, test, and analyze (DMTA). The application of AI is particularly important in the stage of drug synthesis, as it can effectively shorten the cycle of new drug R&D by speeding up the discovery of a new synthetic route for target molecules and reducing the rate of synthetic failure when the structure of the target molecule is known.

7.1. Automated exploration of reaction spaces with AI

In the 1960s, Corey proposed computer-aided synthetic design (CASP) [373] as the earliest AI drug synthesis design. However, due to the lack of computing power at that time, this concept could not be further developed. With the development of ML methods in recent years, CASP has come back into the limelight. CASP mainly consists of three aspects: retrosynthetic planning, reaction condition recommendation, and forward reaction prediction [374]. Retrosynthetic planning, which involves the stepwise splitting of the target molecule into commercially available chemical materials, is an important approach in the design of drug synthesis reactions. MC tree search (MCTS) is a general search technique for sequential decision-making with large branching factors. Segler et al. [375] combined three different neural networks trained with all published reactions with MCTS to predict the best retrosynthetic routes. In comparison with conventional algorithms, the model is 30 times faster and doubles the number of molecules solved.

After designing the synthetic route, the rationality of each step in the synthesis process must also be considered. Researchers have also used AI for the prediction of reaction conditions in order to reduce the time spent on screening reaction conditions. Gao et al. [376] proposed a neural network model to predict appropriate reaction conditions and reaction temperature. They trained the model using ten million examples on Reaxys and tested it on one million reactions outside the training set. Their results showed the model's ability to predict reaction conditions that matched those in the record in 69.6% of those cases. The computational framework DeepReac+ [377] also adopted an active learning strategy to explore the response space more efficiently in order to reduce the time for model learning and prediction.

Forward reaction prediction verifies the feasibility of the designed route by predicting the products. The starting material, which is predicted by retrosynthetic planning, can be replaced by many other compounds, and forward reaction prediction can be used to rank these compounds in order to select the best solution. For example, Coley et al. [378] proposed a neural network model for predicting reaction outcomes. They trained the model with 15 000 reaction examples from the USPTO literature and ranked all the generated candidate compounds to select the product that matched the record. The model used an edit-based representation of the candidate reactions and achieved an accuracy of 71.8%.

In addition to designing new reaction routes based on target molecules, unknown chemical spaces can be explored by synthetic robots based on AI. Recently, a synthetic robot proposed by Granda et al. [26] not only analyzed chemical reactions faster than manual analysis but was also able to predict the reactivity of various reaction combinations on its own and explore the unknown reaction space. The robot model's analysis of samples by nuclear magnetic resonance and infrared spectroscopy is coupled with ML for decision-making, allowing reactions to be evaluated in real time. The outcomes showed that the model can predict the reactivity of about 1000 reaction combinations with over 80% accuracy. Four entirely new reactions were discovered by chemists using real-time data from this robot for prediction. In addition, Caramelli et al. [379] proposed an inexpensive synthetic robot with the ability to network and coordinate multiple reactions in addition to performing chemical reactions autonomously. The robot can also explore new chemical spaces to search for new reaction results and can evaluate the reproducibility of reactions. In conclusion, the invention of intelligent synthesis robots is an important step toward an automated synthesis approach with AI.

7.2. AI usage in automatic drug synthesis

AI-based automated chemical synthesis technologies are freeing researchers from a great deal of manual works by automating experimental processes. Many reactions can already be performed on automated synthetic systems, such as the synthesis of peptides [380], oligonucleotides [381], natural products [382], and various drug molecules [383], as reported earlier. To establish a common standard for automated chemical synthesis, Steiner et al. [35] proposed the Chemputer system and used it to synthesize three drug compounds—diphenhydramine hydrochloride, flufenamide, and sildenafil—in yields comparable to those from manual synthesis. The program they developed, called Chempiler, allows low-level instructions to be compiled in order to synthesize compounds through a modular robotic platform. Moreover, the synthesis process is captured to generate digital code that is shared between platforms, thereby driving the spread of automated chemical synthesis in the laboratory.

In parallel to increasing the automation of reactions, improving the reaction throughput is a goal of automated synthesis, causing high-throughput experiments (HTEs) to receive much attention in recent years. HTEs with 24- or 96-well reactors are capable of performing dozens of reactions in a single experiment [384,385]. In contrast, ultra-high-throughput reactions on the nanoscale can even perform thousands of reactions at a time [386,387]. Of the limited types of reactions that high throughput can currently achieve, heated reactions with homogeneous reactions in low-volatile solvents at room temperature are relatively easy to achieve [388]. Moreover, among the reactions commonly used in THE, metal-catalyzed cross-coupling reactions in which many reaction variables are observed during development are a hot research topic. Ahneman et al. [389] proposed an RF algorithm trained by a high-throughput dataset to predict the tolerance of palladium catalysts to isoxazole during C–N bond formation. The performance of the algorithm was shown to be significantly improved compared with conventional linear regression analysis, and the model was also useful for analyzing the inhibition mechanism of metal catalysts.

As an increasing number of algorithms related to reaction prediction are developed, scientists can identify optimal reaction conditions faster and more accurately, obtain optimal reaction routes, and further explore the reaction space. The integration of these novel and effective algorithms can facilitate the development of automated chemical synthesis platforms, freeing researchers from repetitive tasks [377].

8. Application of AI in other areas related to drug discovery

AI technology has been widely used in the whole process of drug R&D, including target identification, drug design, synthesis, and property evaluation. It has undoubtedly shortened the drug R&D cycle and saved a great deal of experimental cost compared with the traditional experimental process. Scientists are continuing to explore the application of AI technology, as they attempt to use AI in more fields to promote the development of pharmaceutical sciences.

8.1. Facilitating knowledge discovery through literature mining

Every year, numerous papers are published in the fields of medicine, pharmacy, biology, chemistry, materials, and so forth. There is a great deal of relevant expertise in these papers. Mining the literature and linking information with relevant knowledge quickly and purposefully is very important. NLP algorithms can extract the required knowledge from unstructured information in a large number of papers, patents, and published documents. Further analysis of the extracted knowledge can reveal the knowledge associations hidden in many documents and can thereby reduce the workload of researchers in analyzing documents one by one [390]. Long short-term memory (LSTM), gate recurrent units (GRUs), bidirectional encoder representations from transformers (BERT), and transformers, which are commonly used in NLP research, have made their mark in this field [391,392].

MEDLINE is a commonly used corpus in the biomedical field and is an important part of PubMed. For decades, there has been extensive work on text mining this corpus for screening key genes, targets, and drugs and for drug side-effect discovery, drug repositioning, and other research. Researchers have focused on five main areas of text mining in biomedicine—namely, biomedical named entity recognition (NER) and normalization, biomedical text classification, relation extraction (RE), pathway extraction, and hypothesis generation [393]—which has led to many new discoveries. For example, hypothesis generation studies on biomedicine have driven research on drug repositioning [394,395], drug development [396,397], and pharmacovigilance [398,399].

Hundreds of papers are published every day on COVID-19 research, and text mining can be helpful for finding useful knowledge from the vast literature of this research boom. The COVID-19 Open Research Dataset (CORD-19, https://www.semanticscholar.org/cord19) is a corpus containing a large amount of information related to COVID-19, and most text mining models are based on this corpus for information extraction. The COVID-19 text mining model uses NLP correlation models to mine the constructed corpus for the implementation of the following applications: a question–answer (QA) system (to answer questions asked by users, the model system extracts relevant answers from the corpus), a summarization system (for long texts, the main points are automatically inferred to provide users with a quick overview), visualization (the information in the text is visualized to make it easier for users to understand), and others [400]. These findings have greatly helped researchers to cope with the challenge of information overload and to obtain valuable information in a short period of time.

Aside from the examples given above, text mining models driven by DL will have applications in many more scenarios. As time progresses, advances in NLP technology will make it easier for models to understand human language. Then the model will be able to extract knowledge from this unstructured information by relying on contextual associations to extract the focus of the full text. In this way, thousands of related documents will be processed into a knowledge network to provide a rich knowledge base for drug development. For example, the web service—explorer for target significance and novelty (e-TSN) [401])—constructed the world's largest relation map using drug targets and diseases extracted by means of NLP-based text mining. The service aims to visualize target-disease KG and provide approved drugs and associated bioactivity information to assist in prioritizing candidate disease-related proteins. Furthermore, Wang et al. [402] developed a multimodal chemical information reconstruction system (CIRS) that automatically processes, extracts, and aligns heterogeneous structure information from text descriptions and structural images of chemical documents. CIRS is a powerful tool for constructing a structured molecular database based on chemical patents to enrich the near-drug space.

8.2. Advancing the development of precision medicine

Precision medicine usually involves the adoption of different treatment plans for the diseases or symptoms of different people. This approach is the opposite of simplifying (or over-simplifying) the classification method of diseases such that all individuals with

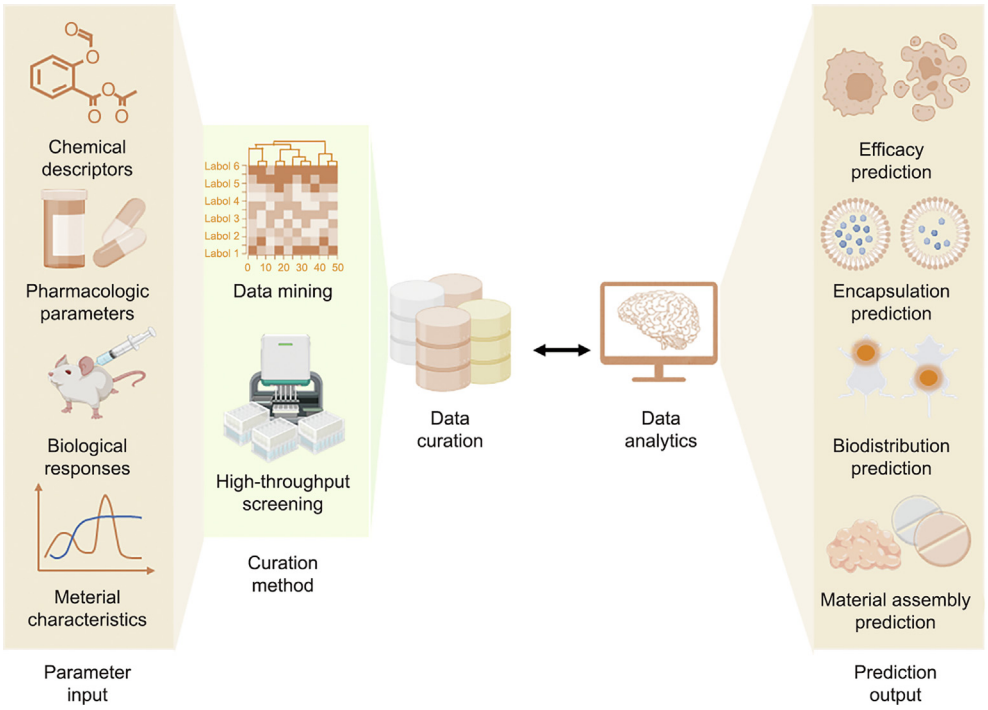

Fig. 4. Using a combination of data curation and ML to enhance the development of nanomedicines.

certain symptoms use the same treatment plan [403]. In society today, the causes of patients' illness are affected by more factors than before, so more accurate diagnosis and treatment plans are required for each patient. The specific concept of precision medicine has been defined as a process [404]. First, information on the patient is needed at different levels, such as the patient's medical history, lifestyle, physical examination results, basic laboratory results, imaging, functional diagnostics, immunology, and omics. This data is then preprocessed to build a relevant model that reflects the patient's situation. Among the data collected, omics data is recognized as the largest and most complex data [404] and has been widely used in the discovery of biomarkers, the identification of disease subgroups, and prognosis prediction [405–408]. In the current era of big data, AI has rapidly advanced the development of precision medicine—especially precision medicine based on omics.

The extensive use of second-generation sequencing technologies has enabled complex diseases to be finely characterized at the molecular scale, especially in the field of tumor research. The global tumor genome sequencing program, represented by the TCGA project, has laid an essential foundation for the molecular typing and precision treatment of tumors. Based on the mRNA expression data of a TCGA dataset through the analysis of differentially expressed genes, Zhao et al. [409] selected the first 40 differentially expressed genes from each type of tumor, merged them to form a feature subset containing 791 different genes, and established a DL model named cancer of unknown primary (CUP)-AI-Dx for predicting the tissue origin and tumor subtype of tumor samples. Yeh et al. [410] studied the transcriptome of patients with severe asthma using the highly variable expressed gene profile of patients' peripheral blood mononuclear cells (PBMCs); their *k*-means clustering analysis of 2048 genes revealed that the genetic characteristics of the transcriptome clusters in patients with asthma determine specific asthma subtypes. In comparison with transcriptomics, the in-depth study of proteomics can help uncover biomarkers and drug targets for different diseases. Rolland et al. [411] used a hierarchical clustering approach to analyze proteomic data from lymphoma patients to reveal specific N-glycoprotein biomarkers in different lymphoma subtypes, thereby providing potential therapeutic targets for precision medicine in lymphoma. Niu et al. [412] identified a combination of protein biomarkers for predicting liver fibrosis, hepatitis, and hepatic steatosis with satisfactory performance using mass spectrometry-based proteomic assays and ML models.

Of course, as mentioned in Section 3, multi-omics technologies are more promising for application than single omics. Many published works explore the molecular mechanisms of disease and the discovery of reliable biomarkers to serve in the diagnosis and treatment of diseases through multi-omics technology. The growing scale of omics data and the increasing development of AI technology will greatly advance the development of precision medicine.

8.3. Utilization of AI in drug formulation and release

With advances in new drug discovery methods, advanced drug delivery systems have expanded rapidly, promoting clinical translation and associated with safety, efficiency, and patient compliance [413,414]. A drug delivery system can be visualized as a "cart" (i.e., a carrier) that transports "goods" (i.e., therapeutics) to the appropriate destination. With the advancement of materials, engineering, and biology technologies, the term "carrier" has expanded to include nanocarriers, cells, eluting devices, and micro–nano robots [415,416]. Compared with conventional drug carriers, nanocarriers can improve drug solubility and mitigate the adverse effects of conventional solubilizers. In addition to protecting the drug from deterioration, nanocarriers can endow the drug with a targeting function [417].

Nevertheless, preparing a suitable nanocarrier is extraordinarily complicated, as it depends on the drugs, excipients, and reaction conditions (including temperature, time, and stirring speed). Experiments alone cannot screen all of these parameters. In addi-

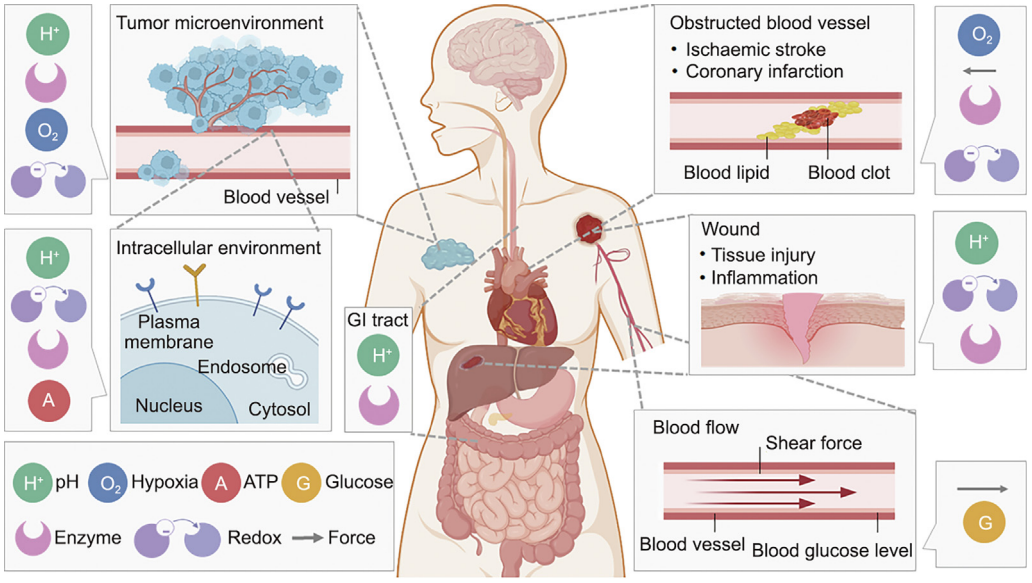

Fig. 5. Bioresponsive design of physiological signal-triggered drug formulations.

tion to determining a drug's molecular target and biological activity [418,419], AI can accurately predict its optimal nano-forming conditions (Fig. 4) [420–422].

Helle et al. [422] predicted particle self-assembly via computational methods. Using quantitative structure-nanoparticle assembly prediction (QSNAP) calculations, they discovered two molecular descriptors for predicting which drugs will form nanoparticles with indocyanine. This method also revealed crucial molecular structural characteristics that permit the self-assembly and the formation of nanoparticles. With the aid of indocyanine sulfate, these drugs were assembled into nanoparticles with a loading efficiency of 90%. The researchers also evaluated the targeted delivery properties of nanoparticles in human colon and primary liver cancer models expressing caveolin 1 (CAV1). Sorafenib- and trametinib-containing nanoparticles were able to selectively target tumors without harming healthy tissue.

In addition, Traverso et al. [9] utilized MD simulations, ML, and a HTE co-aggregation platform to determine which drug-excipient combinations could self-assemble into stable solid drug nanoparticles without additional stabilization. The researchers isolated 100 self-assembled drug nanoparticles from 2.1 million pairs, each containing one of 788 drug candidates and one of 2686 approved excipients. Nanoparticles of sorafenib-glycyrrhizin and terbinafine-taurocholic acid were subjected to proof-of-concept studies *in vitro* and *in vivo*. Both validations suggest that this platform can produce nanoparticles with a high drug loading and enhanced bioavailability, representing a significant step toward personalized drug delivery.

The release pattern of a drug is also crucial for disease treatment. Developing drugs that are released in response to differences in the physiological signals of various organs, tissues, and organelles can enhance the drug's efficacy, prevent toxic and side effects caused by non-specific off-targets, and achieve safe and precise treatment. Multiple endogenous signals—including pH, active redox species, enzymes, glucose, various ions, adenosine triphosphate (ATP), and oxygen—have been incorporated into the design of responsive drug nanocarriers (Fig. 5) [423]. In addition to the material's properties, the target tissue environment influences drug release. AI can facilitate the evaluation of a drug-release mode and can provide feedback for the formulation of drug carriers through ML [424–427].

8.4. Promoting the economic development of the pharmaceutical market

AI has shown itself to be powerful and promising in the pharmaceutical industries, leading to a surge of interest in AI-based drug development from both the scientific and industrial communities. In the past five years, numerous AI-based pharmaceutical companies have been established and have signed collaboration agreements with many large pharmaceutical companies [428]. These shifts have driven massive financing in the drug market, injecting new dynamics into the pharmaceutical economy.

Some of these AI-based pharmaceutical companies focus on a specific stage of the drug discovery pipeline, such as target discovery and the screening of compounds. Some are involved in multiple stages of the pipeline, while others have built end-to-end platforms for new drug discovery [428].

BenevolentAI is a leading AI-based pharmaceutical company that focuses on drug target discovery. Founded in 2013, the company has seen rapid growth in recent years and has emerged as a leader in AI-based drug discovery, attracting significant investor attention. The company was listed in Amsterdam on 6 December 2021 and has a pre-investment valuation of €1.1 billion and a post-investment valuation of up to €1.5 billion. BenevolentAI identifies drug targets for complex diseases through its leading KG technology, which integrates large amounts of publicly available biopharmaceutical data with internal company data. For example, the KG identified baricitinib as a possible treatment for COVID-19 [429]. Through this technology, BenevolentAI has entered into a long-term collaboration with AstraZeneca for target identification in chronic kidney disease, idiopathic pulmonary fibrosis, heart failure, and systemic lupus erythematosus. On 17 May 2022, AstraZeneca made a milestone payment to BenevolentAI for a new target discovery in idiopathic pulmonary fibrosis, which is the third new target identified through BenevolentAI's R&D platform. In addition, BenevolentAI has entered into a new drug discovery collaboration with Johnson & Johnson. The judgment-augmented cognition system (JACS) is a core technology that can focus on processing large amounts of unstructured data in a short period of time through its NLP capabilities. The current market opportunity around AI-led drug discovery capabilities is over 30 billion USD [430].

Table 6
AI-based pharmaceutical companies and their technology platforms.

	Platform	Pharma	Description
Discovery of new targets	JACS	BenevolentAI	A JACS is a judgment–enhancing cognitive system that extracts knowledge from a scattered mass of information and proposes new hypotheses that can be tested, which can identify drug targets by different mechanisms
	XtalCryo™	XtalPi	Combines AI technology with cryo–electron microscopy (EM) structure analysis and kinetic simulation to identify the binding sites of new target proteins
	PandaOmics	Insilico Medicine	Supports multi-omics target discovery and deep biological analysis engine
Drug discovery and design	AtomNet	Atomwise	Learns 3D features of small molecule binding to target targets for lead discovery and optimization
	Ligand Express	Cyclica	Screens and evaluates all protein targets bound to small molecule compounds for drug optimization and repurposing
	Nπφ™	Nuritas	Trains on proprietary experimental data and curated structured and unstructured data to predict novel peptides with on-target efficacy
	Chemistry42	Insilico Medicine	An automated ML platform to rapidly search for new lead-like structures
	CentaurChemist™	Exscientia	Fully automated design of small molecule compounds and calculation of priorities to select the best chemical structure
	Conformetrix	C4x discovery	Measuring dynamic 3D shapes of free drug molecules to accelerate drug design
	XcelaHit™	XtalPi	Ultra-high-throughput VS based on DL method and AI-integrated DNA-encoded library (DEL) technology to rapidly design emerging compounds
Clinical trial design	inClinico	Insilico Medicine	A multimodal data-driven prediction platform for the probability of success (PoS) of a single clinical trial
	Link Recruitment™	LinkDoc	Based on the largest medical big data database in China, relevant data is extracted from clinical trial documents to evaluate the appropriate therapy for patients

In 2019, Insilico Medicine completed a challenge to design new small molecule inhibitors of DDR1 in 21 days using the GENTRL AI system [28]. This challenge caused a great sensation at the time, because it was unimaginable for so many new inhibitors to be discovered in such a short period of time using AI methods. The total time taken was reduced by 1–2 years compared with the traditional process. Insilico Medicine's outstanding performance has made it a hit with investors. In June 2021, Insilico Medicine raised 225 million USD in a Series C round of funding and, in February 2022, it announced the launch of a phase I clinical trial of a small molecule inhibitor for the treatment of idiopathic pulmonary fibrosis [430].

The company Exscientia stands tall in the area of getting small molecules that have been discovered using AI into clinical trials. At a time when AI-based pharmaceutical companies are competing with each other, Exscientia has become the first company to send an AI-discovered drug candidate, DSP-1181, to the clinical stage. This process will take less than 12 months, compared with a historical average of about 4.5 years for this step. In 2021, Exscientia raised a total of approximately 800 million USD through Series C and Series D funding, and an initial public offering (IPO). The company has also raised significant funding through deal partnerships, signing deals with Bristol–Myers Squibb and Sanofi for potential transaction amounts of 1.2 and 5.2 billion USD, respectively. Both deals are focused on drug discovery in the areas of oncology and immunology. Over the decade of Exscientia's development, a complete end-to-end AI drug development pipeline has been progressively established, from target selection to molecular screening and generation. It is this complete pipeline that continues to drive Exscientia's growth. To date, Exscientia has three drugs in the clinical stage, and its market value is highly anticipated upon launch [430].

Thus far, the development of AI-driven drugs is at a historical inflection point, and the average funding for pharmaceutical companies with AI as a core technology has been on the rise. Table 6 provides some information on the core technologies of AI-based pharmaceutical companies. Investors now recognize that drug R&D based on AI technology is becoming a powerful tool to accelerate biopharmaceutical innovation. This technology can provide new insights to accelerate drug discovery by analyzing the biopharmaceutical data that is accumulated and generated on a daily basis. As a result, this field has become a strategic area of focus for pharmaceutical companies and continues to attract capital market attention.

9. Challenges

This review has elaborated on most of the applications of AI in the whole process of drug R&D. However, at the present stage, AI has not really broken down the traditional pharmaceutical system, and many research processes are still waiting for "optimization" by AI. The use of AI for more in-depth research in the field of pharmaceutical preparations is still being gradually explored. For example, some scholars have used AI technology to assist in studying the interaction of drug excipients with biomolecules [431]. In addition to the application areas of AI in the drug development stage that still require expansion, there are limitations in the application of AI to drug discovery.

9.1. Data limitations

The development of AI algorithms cannot be separated from the drive of data. High-quality and accurate data can sometimes enable simple models to outperform complex models. There are many excellent publicly accessible databases for data research, including TTD, CHemBL, Drugbank, CMAP, and PRIDE, but the amount of data is insufficient to support more complex research. The construction of AI algorithms relies heavily on high-quality and sufficient data. The acquisition of high-quality data is a very important issue for sophisticated and complex biological systems, due to the limitations of current technology, and it is costly to process this data into standard data with high confidence. The method, time, and place of operation of each batch of data acquisition are different, making it more difficult to process the acquired data into uniform and valid data [432]. For example, the results obtained by current single-cell RNA-seq (scRNA-seq) vary with their sequencing platforms and often tend to form doublets. Some data is obtained by *in vitro* assays; however, due to the lack of a thorough understanding of the response in the organism, the *in vitro* data often differs significantly from the actual *in vivo* data. Therefore, the prediction results of models trained with the data obtained from *in vitro* experiments are often unconvincing.

These limitations reflect the uneven quality of the data that is currently used. Data imbalance is also a major difficulty in model training. As previously mentioned, positive datasets are readily available in the pharmaceutical field, but negative datasets are often not accurately identified because failed data is often not publicly available. In addition to the problem of data quality and balance, some types of data are generally unavailable to researchers. The key core data for new drug R&D usually originates from drug companies; this part of the data is usually not open source, as drugs are commodities. Similarly, clinical data involves patient privacy and is usually not open source. The problem of data quality and balance requires advances in experimental techniques to obtain more accurate biomedical data in comparison with current data, in order to break the data bottleneck. The development of algorithms such as distributed training can be expected to solve the problem of privacy data to a certain extent. We also appeal to major institutions and companies to disclose as much high-quality data as possible without compromising their own interests.

9.2. Limitations in interpretability

In addition to the limitations of data, DL methods lack interpretability. Compared with traditional ML methods, which often pass through a rigorous mathematical reasoning validation analysis, DL methods are considered to be a black box. Although DL performs better than ML on most tasks, it is often impossible for researchers to understand the reason results of ML are so good. When a DL model yields a new result that contradicts previous research, the lack of interpretability makes the result unacceptable. In particular, compared with other fields, the field of drug discovery has a complete set of knowledge logic, such as the mechanisms of action of molecules, the metabolic mechanisms of molecules, and the regulatory mechanisms of biological pathways. In order to ensure the safety and efficacy of drugs, relevant biological processes must be thoroughly studied, ranging from the physicochemical properties of a drug to what proteins it binds to in the body, how it binds, what biological reactions it triggers, and how it is metabolized. DL can only accept input and give predicted output; it cannot provide sufficient explanations for how this output is derived. For example, for protein function annotation, although DL methods can predict the GOA of a specific protein [70], the computational process is not known and most of the predictions are not accepted when the accuracy is not reliable. Even in terms of data representation methods, no uniform standards have been developed regarding which representation method is more suitable for which study and which representation methods lead to a loss of information.

In the future, the development of DL in the pharmaceutical sciences and industry should focus on improving interpretability as much as possible without compromising accuracy, and should involve the establishment of a set of well-established research methods that combine white-box models with black-box models.

10. Conclusions

In conclusion, AI is advantageous in all aspects of new drug R&D. It can be used in the discovery of drug targets, the design and development of new drugs, preclinical research, clinical trial design, and post-market surveillance to assist in the design of safe and effective drugs, while greatly reducing the cycle time and cost of drug R&D. Some limitations still remain in the AI-based drug R&D process. However, we believe that the emergence of AI is gradually assisting us in unraveling the mystery of large and complex biological systems, and that AI has become an indispensable technology in the drug R&D process. Furthermore, AI technologies will change the R&D paradigm of pharmaceutical sciences in the future, helping us to better overcome complex diseases while providing personalized medicine to patients. In this process, further research is needed to inject new energy into this field.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the Natural Science Foundation of Zhejiang Province (LR21H300001), National Key R&D Program of China (2022YFC3400501), National Natural Science Foundation of China (22220102001, U1909208, 81872798, and 81825020), Leading Talent of the "Ten Thousand Plan"—National High-Level Talents Special Support Plan of China, Fundamental Research Fund of Central University (2018QNA7023), Key R&D Program of Zhejiang Province (2020C03010), "Double Top-Class" University (181201*194232101), Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine), Alibaba–Zhejiang University Joint Research Center of Future Digital Healthcare, and Alibaba Cloud, Information Technology Center of Zhejiang University.

The authors would like to dedicate this article to Prof. Hualiang Jiang, Academician of the Chinese Academy of Sciences (CAS) and Professor in Shanghai Institute of Materia Medica and Lingang Laboratory. Prof. Jiang had devoted great efforts to the cutting-edge research on CADD and artificial intelligence for drug discover (AIDD), and made significant contributions to the development of pharmaceutical sciences. All authors would like to take this opportunity to thank for his kind and persistent supports to their research.

Compliance with ethics guidelines

Mingkun Lu, Jiayi Yin, Qi Zhu, Gaole Lin, Minjie Mou, Fuyao Liu, Ziqi Pan, Nanxin You, Xichen Lian, Fengcheng Li, Hongning Zhang, Lingyan Zheng, Wei Zhang, Hanyu Zhang, Zihao Shen, Zhen Gu, Honglin Li, and Feng Zhu declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nat Rev Drug Discov 2017;16(6):381–2.
[2] Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated costs of pivotal trials for novel therapeutic agents approved by the US food and drug administration, 2015–2016. JAMA Intern Med 2018;178(11):1451–7.
[3] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010;9(3):203–14.
[4] Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008;4(11):682–90.
[5] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10(1):57–63.
[6] Giacomotto J, Ségalat L. High-throughput screening and small animal models, where are we? Br J Pharmacol 2010;160(2):204–16.
[7] Mayr LM, Bojanic D. Novel trends in high-throughput screening. Curr Opin Pharmacol 2009;9(5):580–8.
[8] Shoichet BK. Virtual screening of chemical libraries. Nature 2004;432(7019):862–5.
[9] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004;3(11):935–49.
[10] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

[11] Farabet C, Couprie C, Najman L, Lecun Y. Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1915–29.
[12] Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech 2012;20(1):30–42.
[13] Ding J, Sharon N, Bar-Joseph Z. Temporal modelling using single-cell transcriptomics. Nat Rev Genet 2022;23(6):355–68.
[14] Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000 Res 2016;5(1):182.
[15] Luecken MD, Theis FJ. Current best practices in single-cell RNA-Seq analysis: a tutorial. Mol Syst Biol 2019;15(6):e8746.
[16] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422(6928):198–207.
[17] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021;49(D1): D1388–95.
[18] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–9.
[19] Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 2018;19(2):286–302.
[20] Shi Y, Prieto PL, Zepel T, Grunert S, Hein JE. Automated experimentation powers data science in chemistry. Acc Chem Res 2021;54(3):546–55.
[21] Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat Rev Genet 2021;22(1):3–18.
[22] Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov 2015;14(7):475–86.
[23] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596(7873):590–6.
[24] Ying C, Cai T, Luo S, Zheng S, Ke G, He D, et al. Do transformers really perform badly for graph representation? Adv Neural Inf Process Syst 2021;34 (1):28877–88.
[25] Seyed Tabib NS, Madgwick M, Sudhakar P, Verstockt B, Korcsmaros T, Vermeire S. Big data in IBD: big progress for clinical practice. Gut 2020;69 (8):1520–32.
[26] Granda JM, Donina L, Dragone V, Long DL, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 2018;559(7714):377–81. Corrected in: Nature 2019;570:E67–9.
[27] Zhong F, Xing J, Li X, Liu X, Fu Z, Xiong Z, et al. Artificial intelligence in drug design. Sci China Life Sci 2018;61(10):1191–204.
[28] Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol 2019;37(9):1038–40.
[29] Winter R, Montanari F, Noé F, Clevert DA. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 2019;10(6):1692–701.
[30] Zheng S, Rao J, Song Y, Zhang J, Xiao X, Fang EF, et al. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Brief Bioinform 2021;22(4):bbaa344.
[31] Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Med 2014;6(7):57.
[32] Riniker S, Wang Y, Jenkins JL, Landrum GA. Using information from historical high-throughput screens to predict active compounds. J Chem Inf Model 2014;54(7):1880–91.
[33] Basile AO, Yahi A, Tatonetti NP. Artificial intelligence for drug toxicity and safety. Trends Pharmacol Sci 2019;40(9):624–35.
[34] Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. The SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit. Health 2020;2(10):e549–60.
[35] Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. Science 2019;363(6423):eaav2211.
[36] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69 (Suppl):S36–40.
[37] Zhou Y, Zhang Y, Lian X, Li F, Wang C, Zhu F, et al. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Res 2022;50(D1):D1398–407.
[38] Amahong K, Zhang W, Zhou Y, Zhang S, Yin J, Li F, et al. CovInter: interaction data between coronavirus RNAs and host proteins. Nucleic Acids Res 2022;51 (D1):D546–56.
[39] Liu S, Chen L, Zhang Y, Zhou Y, He Y, Chen Z, et al. M6AREG: m6A-centered regulation of disease development and drug response. Nucleic Acids Res 2022;51(D1):D1333–44.
[40] Sun X, Zhang Y, Li H, Zhou Y, Shi S, Chen Z, et al. DRESIS: the first comprehensive landscape of drug resistance information. Nucleic Acids Res 2022;51(D1). D1263–75.
[41] Wang X, Li F, Qiu W, Xu B, Li Y, Lian X, et al. SYNBIP: synthetic binding proteins for research, diagnosis and therapy. Nucleic Acids Res 2022;50(D1): D560–70.
[42] Zhang S, Sun X, Mou M, Amahong K, Sun H, Zhang W, et al. REGLIV: molecular regulation data of diverse living systems facilitating current multiomics research. Comput Biol Med 2022;148(1):105825.
[43] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 2021;49(D1):D437–51.
[44] Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res 2022;50 (D1):D543–52.
[45] Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 2021;49(D1):D344–54.
[46] Yin J, Sun W, Li F, Hong J, Li X, Zhou Y, et al. VARIDT 1.0: variability of drug transporter database. Nucleic Acids Res 2020;48(D1):D1042–50.
[47] Fu T, Li F, Zhang Y, Yin J, Qiu W, Li X, et al. VARIDT 2.0: structural variability of drug transporter. Nucleic Acids Res 2022;50(D1):D1417–31.
[48] Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res 2022;50(D1):D988–95.
[49] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12(6):996–1006.
[50] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013;41(D1):D991–5.
[51] Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. Nucleic Acids Res 2022;50(D1):D161–4.
[52] Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res 2021;49(D1):D1020–8.
[53] Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res 2020;48(D1):D77–83.
[54] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47(D1): D930–40.
[55] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46 (D1):D1074–82.
[56] Li F, Yin J, Lu M, Mou M, Li Z, Zeng Z, et al. DrugMAP: molecular atlas and pharma-information of all drugs. Nucleic Acids Res 2022;51(D1): D1288–99.
[57] Tang J, Tanoli ZU, Ravikumar B, Alam Z, Rebane A, Vähä-Koskela M, et al. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. Cell Chem Biol 2018;25(2):224–9. .e2.
[58] Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res 2021;49(D1):D1334–46.
[59] Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. Cell 2018;173(2):283–5.
[60] Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET Cytoscape App: exploring and visualizing disease genomics data. Comput Struct Biotechnol J 2021;19 (1):2960–7.
[61] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48(D1):D835–44.
[62] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 2019;47 (D1):D1038–43.
[63] Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology 2020;132(2):379–94.
[64] Cheng W, Ng CA. Using machine learning to classify bioactivity for 3486 per- and polyfluoroalkyl substances (PFASs) from the OECD list. Environ Sci Technol 2019;53(23):13970–80.
[65] Hong J, Luo Y, Mou M, Fu J, Zhang Y, Xue W, et al. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. Brief Bioinform 2020;21 (5):1825–36.
[66] Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. Recent applications of deep learning and machine intelligence on *in silico* drug discovery: methods, tools and databases. Brief Bioinform 2019;20 (5):1878–912.
[67] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 2020;36(2):422–9.
[68] Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics 2018;34(4):660–8.
[69] Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12(1):3168.
[70] Xia W, Zheng L, Fang J, Li F, Zhou Y, Zeng Z, et al. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. Comput Biol Med 2022;145 (1):105465.

[71] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J 2021;19(1):4538–58.
[72] Ubels J, Schaefers T, Punt C, Guchelaar HJ, de Ridder J. RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. Bioinformatics 2020;36(Suppl 2):i601–9.
[73] Yang C, Zhang Y. Delta machine learning to improve scoring-ranking-screening performances of protein-ligand scoring functions. J Chem Inf Model 2022;62(11):2696–712.
[74] Heikamp K, Bajorath J. Support vector machines for drug discovery. Expert Opin Drug Discov 2014;9(1):93–104.
[75] Zhang S, Li X, Zong M, Zhu X, Wang R. Efficient *k*NN classification with different numbers of nearest neighbors. IEEE Trans Neural Netw Learn Syst 2018;29(5):1774–85.
[76] Liu B, He H, Luo H, Zhang T, Jiang J. Artificial intelligence and big data facilitated targeted drug discovery. Stroke Vasc Neurol 2019;4(4):206–13.
[77] Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol 2019;58(1):161–7.
[78] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model 2015;55 (2):263–74.
[79] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35 (5):1285–98.
[80] Hou BJ, Zhou ZH. Learning with interpretable structure from gated RNN. IEEE Trans Neural Netw Learn Syst 2020;31(7):2267–79.
[81] Zhang Z, Chen L, Zhong F, Wang D, Jiang J, Zhang S, et al. Graph neural network approaches for drug-target interactions. Curr Opin Struct Biol 2022;73(1):102327.
[82] Zhang H, Wang Y, Pan Z, Sun X, Mou M, Zhang B, et al. ncRNAInter: a novel strategy based on graph neural network to discover interactions between lncRNA and miRNA. Brief Bioinform 2022;23(6):bbac411.
[83] Sun C, Cao Y, Wei JM, Liu J. Autoencoder-based drug-target interaction prediction by preserving the consistency of chemical properties and functions of drugs. Bioinformatics 2021;37(20):3618–25.
[84] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal 2019;58(1):101552.
[85] Zhou X, Shen F, Liu L, Liu W, Nie L, Yang Y, et al. Graph convolutional network hashing. IEEE Trans Cybern 2020;50(4):1460–72.
[86] Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. AJR Am J Roentgenol 2019;212(1):38–43.
[87] Richards BA, Frankland PW. The persistence and transience of memory. Neuron 2017;94(6):1071–84.
[88] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[89] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.
[90] Sun J, Chen X, Zhang Z, Lai S, Zhao B, Liu H, et al. Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. Sci Rep 2020;10 (1):21122.
[91] Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant Biol 2016;4(4):320–30.
[92] Xue L, Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb Chem High Throughput Screen 2000;3(5):363–72.
[93] Wenzel J, Matter H, Schmidt F. Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. J Chem Inf Model 2019;59(3):1253–68.
[94] Goh GB, Siegel C, Vishnu A, Hodas N. Using rule-based labels for weak supervised learning: a ChemNet for transferable chemical property prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19-23; London UK. New York City: Association for Computing Machinery; 2018. p. 302–10.
[95] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design. Sci Adv 2018;4(7):eaap7885.
[96] Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 2020;12(1):17.
[97] Goh GB, Hodas NO, Siegel C, Vishnu A. Smiles2vec: an interpretable general-purpose deep neural network for predicting chemical properties. 2017. arXiv:171202034.
[98] Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem 2020;63(16):8749–60.
[99] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59 (8):3370–88.
[100] Li K, Xu C, Huang J, Liu W, Zhang L, Wan W, et al. Prediction and identification of the effectors of heterotrimeric G proteins in rice (*Oryza sativa* L.). Brief Bioinform 2017;18(2):270–8.
[101] Wu M, Yang Y, Wang H, Xu Y. A deep learning method to more accurately recall known lysine acetylation sites. BMC Bioinf 2019;20(1):49.
[102] Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, et al. SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. PLoS One 2016;11(8):e0155290.
[103] Zou L, Nan C, Hu F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. Bioinformatics 2013;29 (24):3135–42.
[104] Petrilli P. Classification of protein sequences by their dipeptide composition. Comput Appl Biosci 1993;9(2):205–9.
[105] Seo S, Oh M, Park Y, Kim S. DeepFam: deep learning based alignment-free method for protein family modeling and prediction. Bioinformatics 2018;34 (13):i254–62.
[106] Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics 2017;33(17):2756–8.
[107] Hong J, Luo Y, Zhang Y, Ying J, Xue W, Xie T, et al. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief Bioinform 2020;21(4):1437–47.
[108] Yu CY, Li XX, Yang H, Li YH, Xue WW, Chen YZ, et al. Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. Int J Mol Sci 2018;19(1):183.
[109] Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21(1):10–9.
[110] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43(3):246–55.
[111] Mosier PD, Counterman AE, Jurs PC, Clemmer DE. Prediction of peptide ion collision cross sections from topological molecular structure and amino acid parameters. Anal Chem 2002;74(6):1360–70.
[112] Ren B. Atomic-level-based AI topological descriptors for structure-property correlations. J Chem Inf Comput Sci 2003;43(1):161–9.
[113] Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics 2014;30 (18):2592–7.
[114] Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM. Fast and flexible protein design using deep graph neural networks. Cell Syst 2020;11 (4):402–11. .e4.
[115] Ingraham J, Garg V, Barzilay R, Jaakkola T. Generative models for graph-based protein design. Adv Neural Inf Process Syst 2019;32(1):15820–31.
[116] Greener JG, Moffat L, Jones DT. Design of metalloproteins and novel protein folds using variational autoencoders. Sci Rep 2018;8(1):16189.
[117] Karimi M, Zhu S, Cao Y, Shen Y. *De novo* protein design for novel folds using guided conditional Wasserstein generative adversarial networks. J Chem Inf Model 2020;60(12):5667–81.
[118] Ye Q, Hsieh CY, Yang Z, Kang Y, Chen J, Cao D, et al. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun 2021;12(1):6775.
[119] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 2021;118(15). e2016239118.
[120] Li GB, Yang LL, Wang WJ, Li LL, Yang SY. ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. J Chem Inf Model 2013;53(3):592–600.
[121] Montanari F, Kuhnke L, Ter Laak A, Clevert DA. Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. Molecules 2019;25(1):44.
[122] Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev 2022;55(3):1947–99.
[123] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular *de-novo* design through deep reinforcement learning. J Cheminform 2017;9(1):48.
[124] Dean SN, Alvarez JAE, Zabetakis D, Walper SA, Malanoski AP. PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction. Front Microbiol 2021;12(1):725727.
[125] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res 2021;49(W1):W5–W.
[126] Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform 2021;22(6):bbab159.
[127] Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov 2011;10(7):507–19.
[128] Vincent F, Nueda A, Lee J, Schenone M, Prunotto M, Mercola M. Publisher correction: phenotypic drug discovery: recent successes, lessons learned and new directions. Nat Rev Drug Discov 2022;21(7):541.
[129] Li YH, Li XX, Hong JJ, Wang YX, Fu JB, Yang H, et al. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. Brief Bioinform 2020;21(2):649–62.
[130] Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: tools, advances, and future approaches. J Mol Endocrinol 2019;62(1):21–45.
[131] Fu J, Zhang Y, Wang Y, Zhang H, Liu J, Tang J, et al. Optimization of metabolomic data processing using NOREVA. Nat Protoc 2022;17(1):129–51.
[132] Tang J, Fu J, Wang Y, Li B, Li Y, Yang Q, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Brief Bioinform 2020;21(2):621–36.

[133] Li F, Zhou Y, Zhang Y, Yin J, Qiu Y, Gao J, et al. POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. Brief Bioinform 2022;23(2):bbac040.
[134] Li F, Yin J, Lu M, Yang Q, Zeng Z, Zhang B, et al. ConSIG: consistent discovery of molecular signature from OMIC data. Brief Bioinform 2022;23(4):bbac253.
[135] Yang Q, Li B, Wang P, Xie J, Feng Y, Liu Z, et al. LargeMetabo: an out-of-the-box tool for processing and analyzing large-scale metabolomic data. Brief Bioinform 2022;23(6):bbac455.
[136] Mou M, Pan Z, Lu M, Sun H, Wang Y, Luo Y, et al. Application of machine learning in spatial proteomics. J Chem Inf Model 2022;62(23):5875–95.
[137] Fu J, Yang Q, Luo Y, Zhang S, Tang J, Zhang Y, et al. Label-free proteome quantification and evaluation. Brief Bioinform 2022;24(1):bbac477.
[138] Yang Q, Wang Y, Zhang Y, Li F, Xia W, Zhou Y, et al. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. Nucleic Acids Res 2020;48(W1):W436–48.
[139] Zhang S, Amahong K, Zhang C, Li F, Gao J, Qiu Y, et al. RNA-RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection. Brief Bioinform 2022;23(1):bbab397.
[140] Yang Q, Li B, Tang J, Cui X, Wang Y, Li X, et al. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. Brief Bioinform 2020;21(3):1058–68.
[141] Zhang S, Amahong K, Sun X, Lian X, Liu J, Sun H, et al. The miRNA: a small but powerful RNA for COVID-19. Brief Bioinform 2021;22(2):1137–49.
[142] Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv 2021;49(1):107739.
[143] Network CGAR. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455(7216):1061–8. Corrected in: Nature 2013;494(7438):506.
[144] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 2017;171(6):1437–1452.e17.
[145] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science 2015;347(6220):1260419.
[146] Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature 2014;509(7502):575–81.
[147] Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res 2022;50(D1):D622–31.
[148] Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit 2005;27(6):747–51.
[149] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353–61.
[150] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc database of metabolic pathways and enzymes—a 2019 update. Nucleic Acids Res 2020;48(D1):D445–53.
[151] Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The Reactome pathway knowledgebase 2022. Nucleic Acids Res 2022;50(D1):D687–92.
[152] Zhang Y, Tseng JT, Lien IC, Li F, Wu W, Li H. mRNAsi index: machine learning in mining lung adenocarcinoma stem cell biomarkers. Genes 2020;11(3):257.
[153] Duda M, Zhang H, Li HD, Wall DP, Burmeister M, Guan Y. Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. Transl Psychiatry 2018;8(1):56.
[154] Liu TP, Hsieh YY, Chou CJ, Yang PM. Systematic polypharmacology and drug repurposing via an integrated L1000-based Connectivity Map database mining. R Soc Open Sci 2018;5(11):181321.
[155] Gao Y, Kim S, Lee YI, Lee J. Cellular stress-modulating drugs can potentially be identified by *in silico* screening with Connectivity Map (CMap). Int J Mol Sci 2019;20(22):5601.
[156] Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, et al. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res 2010;38(Suppl 2):W609–14.
[157] Wang X, Shen Y, Wang S, Li S, Zhang W, Liu X, et al. PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. Nucleic Acids Res 2017;45(W1):W356–60.
[158] Wang X, Pan C, Gong J, Liu X, Li H. Enhancing the enrichment of pharmacophore-based target prediction for the polypharmacological profiles of drugs. J Chem Inf Model 2016;56(6):1175–83.
[159] Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, et al. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. Bioinformatics 2013;29(14):1827–9.
[160] Wang X, Chen H, Yang F, Gong J, Li S, Pei J, et al. iDrug: a web-accessible and interactive drug discovery and design platform. J Cheminform 2014;6(1):28.
[161] Noh H, Gunawan R. Inferring gene targets of drugs and chemical compounds from gene expression profiles. Bioinformatics 2016;32(14):2120–7.
[162] Zhu J, Wang J, Wang X, Gao M, Guo B, Gao M, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. Nat Biotechnol 2021;39(11):1444–52.
[163] Woo JH, Shimoni Y, Yang WS, Subramaniam P, Iyer A, Nicoletti P, et al. Elucidating compound mechanism of action by network perturbation analysis. Cell 2015;162(2):441–51.
[164] Li B, Tang J, Yang Q, Li S, Cui X, Li Y, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res 2017;45(W1):W162–70.
[165] CGAR Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474(7353):609–15. Erratum in: Nature 2012;490(7419):292.
[166] Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res 2011;71(13):4550–61.
[167] Fang J, Zhang P, Wang Q, Chiang CW, Zhou Y, Hou Y, et al. Artificial intelligence framework identifies candidate targets for drug repurposing in Alzheimer's disease. Alzheimers Res Ther 2022;14(1):7.
[168] Pabon NA, Xia Y, Estabrooks SK, Ye Z, Herbrand AK, Süß E, et al. Predicting protein targets for drug-like compounds using transcriptomics. PLoS Comput Biol 2018;14(12):e1006651.
[169] Zhong F, Wu X, Yang R, Li X, Wang D, Fu Z, et al. Drug target inference by mining transcriptional data using a novel graph convolutional network framework. Protein Cell 2022;13(4):281–301.
[170] Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell 2019;176(3):535–548.e24.
[171] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods 2018;15(12):1053–8.
[172] Liu F, Li H, Ren C, Bo X, Shu W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Sci Rep 2016;6(1):28517.
[173] Downes DJ, Cross AR, Hua P, Roberts N, Schwessinger R, Cutler AJ, et al. COMBAT Consortium. Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. Nat Genet 2021;53(11):1606–15.
[174] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483(7391):603–7.
[175] Dry JR, Pavey S, Pratilas CA, Harbron C, Runswick S, Hodgson D, et al. Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). Cancer Res 2010;70(6):2264–73.
[176] Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 2019;35(14):i501–9.
[177] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell 2016;166(3):740–54.
[178] Peng W, Chen T, Dai W. Predicting drug response based on multi-omics fusion and graph convolution. IEEE J Biomed Health Inform 2022;26(3):1384–93.
[179] Wang Y, Yang Y, Chen S, Wang J. DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. Brief Bioinform 2021;22(5):bbab048.
[180] Novac N. Challenges and opportunities of drug repositioning. Trends Pharmacol Sci 2013;34(5):267–72.
[181] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature 2009;462(7270):175–81.
[182] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res 2021;49(D1):D1138–43.
[183] Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. Nucleic Acids Res 2022;50(D1):D1282–94.
[184] Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, et al. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res 2021;49(D1):D1160–9.
[185] Urán Landaburu L, Berenstein AJ, Videla S, Maru P, Shanmugam D, Chernomoretz A, et al. TDR Targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration. Nucleic Acids Res 2020;48(D1):D992–D1005.
[186] Chen TF, Chang YC, Hsiao Y, Lee KH, Hsiao YC, Lin YH, et al. DockCoV2: a drug database against SARS-CoV-2. Nucleic Acids Res 2021;49(D1):D1152–9.
[187] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 2021;49(D1):D545–51.
[188] Wang C, Hu G, Wang K, Brylinski M, Xie L, Kurgan L. PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. Bioinformatics 2016;32(4):579–86.
[189] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016;44(D1):D1075–9.
[190] Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open targets platform: supporting systematic drug-target identification and prioritisation. Nucleic Acids Res 2021;49(D1):D1302–10.
[191] Gao Z, Li H, Zhang H, Liu X, Kang L, Luo X, et al. PDTD: a web-accessible protein database for drug target identification. BMC Bioinf 2008;9(1):104.

[192] RDKit: open-source cheminformatics software [Internet]. Basel: T5 Informatics GmbH; [Accessed 9 Feb 2023]. Available from: https://www.rdkit.org/.
[193] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open babel: an open chemical toolbox. J Cheminform 2011;3(1):33.
[194] Daylight Toolkit: C-language interface for SMILES™, SMARTS®, and SMIRKS® [Internet]. Laguna Niguel: Daylight Chemical Information Systems, Inc.; [Accessed 9 Feb 2023]. Available from: https://www.daylight.com/products/toolkit.html.
[195] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 2003;43(2):493–500.
[196] Toolkits OpenEye. Internet. Santa Fe: OpenEye Scientific Software, Inc.; 2022.2.2 [Accessed 9 Feb 2023]. Available from:.
[197] Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. ChemmineR: a compound mining framework for R. Bioinformatics 2008;24(15):1733–4.
[198] Indigo Toolkit [Internet]. Newtown: EPAM System, Inc.; [Accessed 9 Feb 2023]. Available from: https://lifescience.opensource.epam.com/indigo/.
[199] Liu X, Jiang H, Li H. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. J Chem Inf Model 2011;51(9):2372–85.
[200] Lu W, Liu X, Cao X, Xue M, Liu K, Zhao Z, et al. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 2. Prospective case study in the discovery of diverse p90 ribosomal S6 protein kinase 2 inhibitors to suppress cell migration. J Med Chem 2011;54(10):3564–74.
[201] He G, Song Y, Wei W, Wang X, Lu X, Li H. eSHAFTS: integrated and graphical drug design software based on 3D molecular similarity. J Comput Chem 2019;40(6):826–38.
[202] Zhang P, Tao L, Zeng X, Qin C, Chen S, Zhu F, et al. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. Brief Bioinform 2017;18(6):1057–70.
[203] Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res 2013;41(W1):W29–33.
[204] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22(22):4673–80.
[205] Holm L, Laakso LM. Dali server update. Nucleic Acids Res 2016;44(W1):W351–5.
[206] Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins 2004;56(1):143–56.
[207] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9.
[208] Li S, Cai C, Gong J, Liu X, Li H. A fast protein binding site comparison algorithm for proteome-wide protein function prediction and drug repurposing. Proteins 2021;89(11):1541–56.
[209] Prlić A, Bliven S, Rose PW, Bluhm WF, Bizon C, Godzik A, et al. Pre-calculated protein structure alignments at the RCSB PDB website. Bioinformatics 2010;26(23):2983–5.
[210] Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339(3):607–33.
[211] Gao M, Skolnick J. APoc: large-scale identification of similar protein pockets. Bioinformatics 2013;29(5):597–604.
[212] Brylinski M. eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models. PLoS Comput Biol 2014;10(9):e1003829.
[213] Björkholm P, Daniluk P, Kryshtafovych A, Fidelis K, Andersson R, Hvidsten TR. Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. Bioinformatics 2009;25(10):1264–70.
[214] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404–5.
[215] Nayal M, Honig B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. Proteins 2006;63(4):892–906.
[216] Cao DS, Liu S, Xu QS, Lu HM, Huang JH, Hu QN, et al. Large-scale prediction of drug-target interactions using protein sequences and drug topological structures. Anal Chim Acta 2012;752(1):1–10.
[217] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 2018;34(17):i821–9.
[218] Rayhan F, Ahmed S, Shatabda S, Farid DM, Mousavian Z, Dehzangi A, et al. iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. Sci Rep 2017;7(1):17731.
[219] Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug-target interaction prediction. Bioinformatics 2021;36(22–23):5545–7.
[220] Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 2009;25(18):2397–403.
[221] 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(13):i232–40.
[222] Yıldırım MA, Goh KI, Cusick ME, Barabási AL, Vidal M. Drug-target network. Nat Biotechnol 2007;25(10):1119–26.
[223] Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 2017;8(1):573.
[224] Zeng X, Zhu S, Lu W, Liu Z, Huang J, Zhou Y, et al. Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci 2020;11(7):1775–97.
[225] Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models. Brief Bioinform 2021;22(2):1679–93.
[226] Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R. Combining drug and gene similarity measures for drug-target elucidation. J Comput Biol 2011;18(2):133–45.
[227] Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I. Predicting drug-target interactions using probabilistic matrix factorization. J Chem Inf Model 2013;53(12):3399–409.
[228] Sydow D, Burggraaff L, Szengel A, van Vlijmen HWT, IJzerman AP, van Westen GJP,, et al. Advances and challenges in computational target prediction. J Chem Inf Model 2019;59(5):1728–42.
[229] Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. Brief Bioinform 2021;22(1):247–69.
[230] Zhang X, Li L, Ng MK, Zhang S. Drug-target interaction prediction by integrating multiview network data. Comput Biol Chem 2017;69(1):185–93.
[231] Zhang W, Chen Y, Li D. Drug-target interaction prediction through label propagation with linear neighborhood information. Molecules 2017;22(12):2056.
[232] van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PLoS One 2013;8(6):e66952.
[233] He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J Cheminform 2017;9(1):24.
[234] Sharma A, Rani R. BE-DTI': ensemble framework for drug target interaction prediction using dimensionality reduction and active learning. Comput Methods Programs Biomed 2018;165(1):151–62.
[235] Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput Biol 2016;12(2):e1004760.
[236] Bolgár B, Antal P. VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization. BMC Bioinf 2017;18(1):440.
[237] Li L, Cai M. Drug target prediction by multi-view low rank embedding. IEEE/ACM Trans Comput Biol Bioinform 2019;16(5):1712–21.
[238] Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 2012;8(5):e1002503.
[239] Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst 2012;8(7):1970–8.
[240] Chen H, Zhang Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. PLoS One 2013;8(5):e62975.
[241] Alaimo S, Pulvirenti A, Giugno R, Ferro A. Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics 2013;29(16):2004–8.
[242] Mongia A, Majumdar A. Drug-target interaction prediction using multi graph regularized nuclear norm minimization. PLoS One 2020;15(1):e0226484.
[243] Wang Y, Zeng J. Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics 2013;29(13):i126–34.
[244] Shi H, Liu S, Chen J, Li X, Ma Q, Yu B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics 2019;111(6):1839–52.
[245] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-learning-based drug-target interaction prediction. J Proteome Res 2017;16(4):1401–9.
[246] 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(6):e1007129.
[247] Xie L, He S, Song X, Bo X, Zhang Z. Deep learning-based transcriptome data classification for drug-target interaction prediction. BMC Genomics 2018;19(Suppl 7):667.
[248] Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design—a review. Curr Top Med Chem 2010;10(1):95–115.
[249] Jing Y, Bian Y, Hu Z, Wang L, Xie XQ. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J 2018;20(3):58. Corrected in: AAPS J 2018;20(4):79.
[250] Hessler G, Baringhaus KH. Artificial intelligence in drug design. Molecules 2018;23(10):2520.
[251] Burello E, Worth AP. QSAR modeling of nanomaterials. Wiley Interdiscip Rev Nanomed Nanobiotechnol 2011;3(3):298–306.
[252] Xue W, Fu T, Deng S, Yang F, Yang J, Zhu F. Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram. ACS Chem Neurosci 2022;13(3):340–51.
[253] Ballante F, Kooistra AJ, Kampen S, de Graaf C, Carlsson J. Structure-based virtual screening for ligands of G protein-coupled receptors: what can molecular docking do for you? Pharmacol Rev 2021;73(4):1698–736.
[254] Shin WH, Zhu X, Bures MG, Kihara D. Three-dimensional compound comparison methods and their application in drug discovery. Molecules 2015;20(7):12841–62.

[255] Ghislat G, Rahman T, Ballester PJ. Recent progress on the prospective application of machine learning to structure-based virtual screening. Curr Opin Chem Biol 2021;65(1):28–34.
[256] Liu M, Wang S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. J Comput Aided Mol Des 1999;13(5):435–51.
[257] Sneha P, George Priya Doss C. Molecular dynamics: new frontier in personalized medicine. Adv Protein Chem Struct Biol 2016;102(1):181–224.
[258] Xue W, Yang F, Wang P, Zheng G, Chen Y, Yao X, et al. What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. ACS Chem Neurosci 2018;9(5):1128–40.
[259] Xie QQ, Zhong L, Pan YL, Wang XY, Zhou JP, Di-Wu L, et al. Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met. Eur J Med Chem 2011;46(9):3675–80.
[260] Pereira JC, Caffarena ER, Dos Santos CN. Boosting docking-based virtual screening with deep learning. J Chem Inf Model 2016;56(12):2495–506.
[261] Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem 2006;49(23):6789–801.
[262] Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: combining techniques to model RNA—small molecule complexes. RNA 2009;15(6):1219–30.
[263] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31(2):455–61.
[264] AbdulHameed MD, Ippolito DL, Wallqvist A. Predicting rat and human pregnane X receptor activators using Bayesian classification models. Chem Res Toxicol 2016;29(10):1729–40.
[265] Martin EJ, Polyakov VR, Tian L, Perez RC. Profile-QSAR 2.0: kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds. J Chem Inf Model 2017;57(8):2077–88.
[266] Chen JJF, Visco Jr DP. Developing an *in silico* pipeline for faster drug candidate discovery: virtual high throughput screening with the signature molecular descriptor using support vector machine models. Chem Eng Sci 2017;159(1):31–42.
[267] Myint KZ, Wang L, Tong Q, Xie XQ. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol Pharm 2012;9(10):2912–23.
[268] Jaén-Oltra J, Salabert-Salvador MT, García-March FJ, Pérez-Giménez F, Tomás-Vert F. Artificial neural network applied to prediction of fluorquinolone antibacterial activity by topological methods. J Med Chem 2000;43(6):1143–8.
[269] Lenselink EB, Ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform 2017;9(1):45.
[270] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol 2007;25(2):197–206.
[271] Xiao T, Qi X, Chen Y, Jiang Y. Development of ligand-based big data deep neural network models for virtual screening of large compound libraries. Mol Inform 2018;37(11):1800031.
[272] Fang J, Wang L, Li Y, Lian W, Pang X, Wang H, et al. AlzhCPI: a knowledge base for predicting chemical-protein interactions towards Alzheimer's disease. PLoS One 2017;12(5):e0178347.
[273] Bender A, Mussa HY, Glen RC. Screening for dihydrofolate reductase inhibitors using MOLPRINT 2D, a fast fragment-based method employing the naïve Bayesian classifier: limitations of the descriptor and the importance of balanced chemistry in training and test sets. J Biomol Screen 2005;10(7):658–66.
[274] Abdo A, Chen B, Mueller C, Salim N, Willett P. Ligand-based virtual screening using Bayesian networks. J Chem Inf Model 2010;50(6):1012–20.
[275] Li Y, Wang L, Liu Z, Li C, Xu J, Gu Q, et al. Predicting selective liver X receptor β agonists using multiple machine learning methods. Mol Biosyst 2015;11(5):1241–50.
[276] Fang J, Yang R, Gao L, Zhou D, Yang S, Liu AL, et al. Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery. J Chem Inf Model 2013;53(11):3009–20.
[277] Schneider P, Tanrikulu Y, Schneider G. Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. Curr Med Chem 2009;16(3):258–66.
[278] Hristozov D, Oprea TI, Gasteiger J. Ligand-based virtual screening by novelty detection with self-organizing maps. J Chem Inf Model 2007;47(6):2044–62.
[279] Reker D, Rodrigues T, Schneider P, Schneider G. Identifying the macromolecular targets of *de novo*-designed chemical entities through self-organizing map consensus. Proc Natl Acad Sci USA 2014;111(11):4067–72.
[280] Stojanović L, Popović M, Tijanić N, Rakočević G, Kalinić M. Improved scaffold hopping in ligand-based virtual screening using neural representation learning. J Chem Inf Model 2020;60(10):4629–39.
[281] Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, et al. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget 2017;8(7):10883–90.
[282] Xu Y, Chen P, Lin X, Yao H, Lin K. Discovery of CDK4 inhibitors by convolutional neural networks. Future Med Chem 2019;11(3):165–77.
[283] Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. ACS Cent Sci 2017;3(4):283–93.
[284] Zhou Z, Kearnes S, Li L, Zare RN, Riley P. Optimization of molecules via deep reinforcement learning. Sci Rep 2019;9(1):10752. Corrected in: Sci Rep 2020;10(1):10478.
[285] Hartenfeller M, Schneider G. *De novo* drug design. Methods Mol Biol 2011;672(1):299–323.
[286] Segall M. Advances in multiparameter optimization methods for *de novo* drug design. Expert Opin Drug Discov 2014;9(7):803–17.
[287] Schneider G, Fechner U. Computer-based *de novo* design of drug-like molecules. Nat Rev Drug Discov 2005;4(8):649–63.
[288] Sohn JI, Nam JW. The present and future of *de novo* whole-genome assembly. Brief Bioinform 2018;19(1):23–40.
[289] Schneider G, Clark DE. Automated *de novo* drug design: are we nearly there yet? Angew Chem Int Ed Engl 2019;58(32):10792–803.
[290] Xiong J, Xiong Z, Chen K, Jiang H, Zheng M. Graph neural networks for automated *de novo* drug design. Drug Discov Today 2021;26(6):1382–93.
[291] Pereira T, Abbasi M, Ribeiro B, Arrais JP. Diversity oriented deep reinforcement learning for targeted molecule generation. J Cheminform 2021;13(1):21.
[292] Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J. Deep reinforcement learning for multiparameter optimization in *de novo* drug design. J Chem Inf Model 2019;59(7):3166–76.
[293] Maziarka Ł, Pocha A, Kaczmarczyk J, Rataj K, Danel T, Warchoł M. Mol-CycleGAN: a generative model for molecular optimization. J Cheminform 2020;12(1):2.
[294] 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). ChemRxiv. Cambridge: Cambridge Open Engage; 2017.
[295] Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, et al. Reinforced adversarial neural computer for *de novo* molecular design. J Chem Inf Model 2018;58(6):1194–204.
[296] Harel S, Radinsky K. Prototype-based compound discovery using deep generative models. Mol Pharm 2018;15(10):4406–16.
[297] Wilman W, Wrobel S, Bielska W, Deszynski P, Dudzic P, Jaszczyszyn I, et al. Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. Brief Bioinform 2022;23(4):bbac267.
[298] Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning. Patterns 2022;3(2):100406.
[299] Sivasubramanian A, Sircar A, Chaudhury S, Gray JJ. Toward high-resolution homology modeling of antibody F_v regions and application to antibody-antigen docking. Proteins 2009;74(2):497–514.
[300] Schneider C, Buchanan A, Taddese B, Deane CM. DLAB: deep learning methods for structure-based virtual screening of antibodies. Bioinformatics 2022;38(2):377–83.
[301] Eguchi RR, Choe CA, Huang PS. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. PLoS Comput Biol 2022;18(6):e1010271.
[302] Raybould MIJ, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, et al. Five computational developability guidelines for therapeutic antibody profiling. Proc Natl Acad Sci USA 2019;116(10):4025–30.
[303] Kim JH, Hong HJ. Humanization by CDR grafting and specificity-determining residue grafting. Methods Mol Biol 2012;907(1):237–45.
[304] Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. Patterns 2022;3(7):100513.
[305] Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinform Adv 2022;2(1):vbac046.
[306] Fu J, Zhang Y, Liu J, Lian X, Tang J, Zhu F. Pharmacometabonomics: data processing and statistical analysis. Brief Bioinform 2021;22(5):bbab138.
[307] Meanwell NA. Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety. Chem Res Toxicol 2011;24(9):1420–56.
[308] Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 2004;1(4):337–41.
[309] Zhang MQ, Wilkinson B. Drug discovery beyond the 'rule-of-five'. Curr Opin Biotechnol 2007;18(6):478–88.
[310] Manallack DT, Prankerd RJ, Yuriev E, Oprea TI, Chalmers DK. The significance of acid/base properties in drug discovery. Chem Soc Rev 2013;42(2):485–96.
[311] Zhang H, Xiang ML, Ma CY, Huang Q, Li W, Xie Y, et al. Three-class classification models of logS and logP derived by using GA-CG-SVM approach. Mol Divers 2009;13(2):261–8.
[312] Jorgensen WL, Duffy EM. Prediction of drug solubility from Monte Carlo simulations. Bioorg Med Chem Lett 2000;10(11):1155–8.
[313] Kamlet MJ, Doherty RM, Abboud JL, Abraham MH, Taft RW. Linear solvation energy relationships: 36. molecular properties governing solubilities of organic nonelectrolytes in water. J Pharm Sci 1986;75(4):338–49.
[314] Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev 2019;119(18):10520–94.

[315] Elder D, Holm R. Aqueous solubility: simple predictive methods (*in silico*, *in vitro* and bio-relevant approaches). Int J Pharm 2013;453(1):3–11.
[316] Hewitt M, Cronin MT, Enoch SJ, Madden JC, Roberts DW, Dearden JC. *In silico* prediction of aqueous solubility: the solubility challenge. J Chem Inf Model 2009;49(11):2572–87.
[317] Francoeur PG, Koes DR. SolTranNet-a machine learning tool for fast aqueous solubility prediction. J Chem Inf Model 2021;61(6):2530–6.
[318] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[319] Shen WX, Zeng X, Zhu F, Wang YL, Qin C, Tan Y, et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. Nat Mach Intell 2021;3(4):334–43.
[320] Yin J, Li F, Zhou Y, Mou M, Lu Y, Chen K, et al. INTEDE: interactome of drug-metabolizing enzymes. Nucleic Acids Res 2021;49(D1):D1233–43.
[321] Cheng F, Li W, Liu G, Tang Y. *In silico* ADMET prediction: recent advances, current challenges and future trends. Curr Top Med Chem 2013;13(11):1273–89.
[322] Wang Y, Xing J, Xu Y, Zhou N, Peng J, Xiong Z, et al. *In silico* ADME/T modelling for rational drug design. Q Rev Biophys 2015;48(4):488–515.
[323] Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. Drug Discov Today 2019;24(5):1157–65.
[324] Tao L, Zhang P, Qin C, Chen SY, Zhang C, Chen Z, et al. Recent progresses in the exploration of machine learning methods as *in-silico* ADME prediction tools. Adv Drug Deliv Rev 2015;86(1):83–100.
[325] Rácz A, Bajusz D, Miranda-Quintana RA, Héberger K. Machine learning models for classification tasks related to drug safety. Mol Divers 2021;25(3):1409–24.
[326] Vandenberg JI, Perry MD, Perrin MJ, Mann SA, Ke Y, Hill AP. hERG K^+ channels: structure, function, and clinical significance. Physiol Rev 2012;92(3):1393–478.
[327] Kaisar MA, Sajja RK, Prasad S, Abhyankar VV, Liles T, Cucullo L. New experimental models of the blood-brain barrier for CNS drug discovery. Expert Opin Drug Discov 2017;12(1):89–103.
[328] Smyth MJ, Krasovskis E, Sutton VR, Johnstone RW. The drug efflux protein, P-glycoprotein, additionally protects drug-resistant tumor cells from multiple forms of caspase-dependent apoptosis. Proc Natl Acad Sci USA 1998;95(12):7024–9.
[329] Rácz A, Keserű GM. Large-scale evaluation of cytochrome P450 2C9 mediated drug interaction potential with machine learning-based consensus modeling. J Comput Aided Mol Des 2020;34(8):831–9.
[330] 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(1):30.
[331] Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC Med 2016;14(1):10. Corrected in: BMC Med 2019;17(1):56.
[332] Alves VM, Golbraikh A, Capuzzi SJ, Liu K, Lam WI, Korn DR, et al. Multi-descriptor read across (MuDRA): a simple and transparent approach for developing accurate quantitative structure-activity relationship models. J Chem Inf Model 2018;58(6):1214–23.
[333] Lei T, Chen F, Liu H, Sun H, Kang Y, Li D, et al. ADMET evaluation in drug discovery. Part 17: development of quantitative and qualitative prediction models for chemical-induced respiratory toxicity. Mol Pharm 2017;14(7):2407–21.
[334] Zhu L, Zhao J, Zhang Y, Zhou W, Yin L, Wang Y, et al. ADME properties evaluation in drug discovery: *in silico* prediction of blood-brain partitioning. Mol Divers 2018;22(4):979–90.
[335] Su BH, Tu YS, Lin C, Shao CY, Lin OA, Tseng YJ. Rule-based prediction models of cytochrome P450 inhibition. J Chem Inf Model 2015;55(7):1426–34.
[336] Yang M, Chen J, Xu L, Shi X, Zhou X, Xi Z, et al. A novel adaptive ensemble classification framework for ADME prediction. RSC Adv 2018;8(21):11661–83.
[337] Radchenko EV, Dyabina AS, Palyulin VA. Towards deep neural network models for the prediction of the blood-brain barrier permeability for diverse organic compounds. Molecules 2020;25(24):5901.
[338] Wang D, Liu W, Shen Z, Jiang L, Wang J, Li S, et al. Deep learning based drug metabolites prediction. Front Pharmacol 2020;10(1):1586.
[339] Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, et al. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. Bioinformatics 2019;35(6):1067–9.
[340] Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 2017;7(1):42717.
[341] Banerjee P, Eckert AO, Schrey AK, Preissner R. ProTox-II: a webserver for the prediction of toxicity of chemicals. Nucleic Acids Res 2018;46(W1):W257–63.
[342] Pires DE, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 2015;58(9):4066–72.
[343] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol 2014;32(1):40–51.
[344] Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019;40(8):577–91.
[345] Perez-Gracia JL, Sanmamed MF, Bosch A, Patiño-Garcia A, Schalper KA, Segura V, et al. Strategies to design clinical studies to identify predictive biomarkers in cancer research. Cancer Treat Rev 2017;53(1):79–97.
[346] Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018;1(1):53–68.
[347] Palmqvist S, Insel PS, Zetterberg H, Blennow K, Brix B, Stomrud E, et al. Alzheimer's Disease Neuroimaging Initiative, Swedish BioFINDER Study. Accurate risk estimation of β-amyloid positivity to identify prodromal Alzheimer's disease: cross-validation study of practical algorithms. Alzheimers Dement 2019;15(2):194–204.
[348] Romero K, Ito K, Rogers JA, Polhamus D, Qiu R, Stephenson D, et al. Alzheimer's Disease Neuroimaging Initiative for the Coalition Against Major Diseases. The future is now: model-based clinical trial design for Alzheimer's disease. Clin Pharmacol Ther 2015;97(3):210–4.
[349] Bain EE, Shafner L, Walling DP, Othman AA, Chuang-Stein C, Hinkle J, et al. Use of a novel artificial intelligence platform on mobile devices to assess dosing compliance in a phase 2 clinical trial in subjects with schizophrenia. JMIR Mhealth Uhealth 2017;5(2):e18.
[350] Yauney G, Shah P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. In: Proceedings of the 3rd Machine Learning for Healthcare Conference; 2018 Aug 17-18; Stanford, CA, USA; 2018. p. 161-226.
[351] Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 1995;51(1):228–35.
[352] Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf 2013;36(Suppl 1):143–58.
[353] Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov 2010;20(3):361–87.
[354] Morel M, Bacry E, Gaïffas S, Guilloux A, Leroy F. ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection. Biostatistics 2020;21(4):758–74.
[355] Ben Abacha A, Chowdhury MFM, Karanasiou A, Mrabet Y, Lavelli A, Zweigenbaum P. Text mining for pharmacovigilance: using machine learning for drug name recognition and drug-drug interaction extraction and classification. J Biomed Inform 2015;58(1):122–32.
[356] Mower J, Subramanian D, Cohen T. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. J Am Med Inform Assoc 2018;25(10):1339–50.
[357] Lorberbaum T, Nasir M, Keiser MJ, Vilar S, Hripcsak G, Tatonetti NP. Systems pharmacology augments drug safety surveillance. Clin Pharmacol Ther 2015;97(2):151–8.
[358] Enshaei A, Robson CN, Edmondson RJ. Artificial intelligence systems as prognostic and predictive tools in ovarian cancer. Ann Surg Oncol 2015;22(12):3970–5.
[359] Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM Trans Comput Biol Bioinform 2018;16(3):841–50.
[360] Chi CL, Street WN, Wolberg WH. Application of artificial neural network-based survival analysis on two breast cancer datasets. AMIA Annu Symp Proc 2007;2007(1):130–4.
[361] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34(2):113–27.
[362] Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics 2007;23(1):30–7.
[363] Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006;22(14):e184–90.
[364] Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform 2017;108(1):1–8.
[365] Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7(1):12474.
[366] Biglarian A, Hajizadeh E, Kazemnejad A, Zali M. Application of artificial neural network in predicting the survival rate of gastric cancer patients. Iran J Public Health 2011;40(2):80–6.
[367] Zhu Y, Wang QC, Xu MD, Zhang Z, Cheng J, Zhong YS, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest Endosc 2019;89(4):806–815.e1.
[368] Zhu L, Luo W, Su M, Wei H, Wei J, Zhang X, et al. Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients. Biomed Rep 2013;1(5):757–60.
[369] Tian DW, Wu ZL, Jiang LM, Gao J, Wu CL, Hu HL. Neural precursor cell expressed, developmentally downregulated 8 promotes tumor progression and predicts poor prognosis of patients with bladder cancer. Cancer Sci 2019;110(1):458–67.

[370] Hasnain Z, Mason J, Gill K, Miranda G, Gill IS, Kuhn P, et al. Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients. PLoS One 2019;14(2):e0210976.
[371] Kuo RJ, Huang MH, Cheng WC, Lin CC, Wu YH. Application of a two-stage fuzzy neural network to a prostate cancer prognosis system. Artif Intell Med 2015;63(2):119–33.
[372] Zhang S, Xu Y, Hui X, Yang F, Hu Y, Shao J, et al. Improvement in prediction of prostate cancer prognosis with somatic mutational signatures. J Cancer 2017;8(16):3261–7.
[373] Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. Science 1969;166(3902):178–92.
[374] 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(16):8667–82.
[375] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555(7698):604–10.
[376] Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci 2018;4(11):1465–76.
[377] Gong Y, Xue D, Chuai G, Yu J, Liu Q. DeepReac+: deep active learning for quantitative modeling of organic chemical reactions. Chem Sci 2021;12 (43):14459–72.
[378] Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. ACS Cent Sci 2017;3(5):434–43.
[379] Caramelli D, Salley D, Henson A, Camarasa GA, Sharabi S, Keenan G, et al. Networking chemical robots for reaction multitasking. Nat Commun 2018;9 (1):3406.
[380] Merrifield RB. Automated synthesis of peptides. Science 1965;150 (3693):178–85.
[381] Alvarado-Urbina G, Sathe GM, Liu WC, Gillen MF, Duck PD, Bender R, et al. Automated synthesis of gene fragments. Science 1981;214(4518):270–4.
[382] Doi T, Fuse S, Miyamoto S, Nakai K, Sasuga D, Takahashi T. A formal total synthesis of taxol aided by an automated synthesizer. Chem Asian J 2006;1 (3):370–83.
[383] Boström J, Brown DG, Young RJ, Keserü GM. Expanding the medicinal chemistry synthetic toolbox. Nat Rev Drug Discov 2018;17(10):709–27. Erratum in: Nat Rev Drug Discov 2018;17(12):922.
[384] Bellomo A, Celebi-Olcum N, Bu X, Rivera N, Ruck RT, Welch CJ, et al. Rapid catalyst identification for the synthesis of the pyrimidinone core of HIV integrase inhibitors. Angew Chem Int Ed Engl 2012;51(28):6912–5.
[385] Dreher SD, Dormer PG, Sandrock DL, Molander GA. Efficient cross-coupling of secondary alkyltrifluoroborates with aryl chlorides—reaction discovery using parallel microscale experimentation. J Am Chem Soc 2008;130(29):9257–9.
[386] Buitrago Santanilla A, Regalado EL, Pereira T, Shevlin M, Bateman K, Campeau LC, et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. Science 2015;347(6217):49–53.
[387] Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, et al. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science 2018;359(6374):429–34.
[388] Shevlin M. Practical high-throughput experimentation for chemists. ACS Med Chem Lett 2017;8(6):601–7.
[389] Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C-N cross-coupling using machine learning. Science 2018;360(6385):186–90.
[390] Isayev O. Text mining facilitates materials discovery. Nature 2019;571 (7763):42–3.
[391] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[392] Sun C, Yang Z, Luo L, Wang L, Wang J. A deep learning approach with deep contextualized word representations for chemical-protein interaction extraction from biomedical literature. IEEE Access 2019;7(1):151034–46.
[393] Zhao S, Su C, Lu Z, Wang F. Recent advances in biomedical literature mining. Brief Bioinform 2021;22(3):bbaa057.
[394] Deftereos SN, Andronis C, Friedla EJ, Persidis A, Persidis A. Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdiscip Rev Syst Biol Med 2011;3(3):323–34.
[395] Yang HT, Ju JH, Wong YT, Shmulevich I, Chiang JH. Literature-based discovery of new candidates for drug repurposing. Brief Bioinform 2017;18(3):488–97.
[396] Zhang R, Cairelli MJ, Fiszman M, Kilicoglu H, Rindflesch TC, Pakhomov SV, et al. Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs. Cancer Inform 2014;13(Suppl 1):103–11.
[397] Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, et al. Analysis of genomic and proteomic data using advanced literature mining. J Proteome Res 2003;2(4):405–12.
[398] Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. J Biomed Inform 2014;52(1):293–310.
[399] Malec SA, Wei P, Bernstam EV, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. J Biomed Inform 2021;117(1):103719.
[400] Wang LL, Lo K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. Brief Bioinform 2021;22(2):781–99.
[401] Feng Z, Shen Z, Li H, Li S. e-TSN: an interactive visual exploration platform for target-disease knowledge mapping from literature. Brief Bioinform 2022;23 (6):bbac465.
[402] Wang J, Shen Z, Liao Y, Yuan Z, Li S, He G, et al. Multi-modal chemical information reconstruction from images and texts for exploring the near-drug space. Brief Bioinform 2022;23(6):bbac461.
[403] Ahn AC, Tewari M, Poon CS, Phillips RS. The limits of reductionism in medicine: could systems biology offer an alternative? PLoS Med 2006;3(6): e208.
[404] König IR, Fuchs O, Hansen G, von Mutius E, Kopp MV. What is precision medicine? Eur Respir J 2017;50(4):1700391.
[405] Antman EM, Loscalzo J. Precision medicine in cardiology. Nat Rev Cardiol 2016;13(10):591–602.
[406] Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013;346:e5793.
[407] Tang J, Mou M, Wang Y, Luo Y, Zhu F. MetaFS: performance assessment of biomarker discovery in metaproteomics. Brief Bioinform 2021;22(3):bbaa105.
[408] Yang Q, Li B, Chen S, Tang J, Li Y, Li Y, et al. MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. J Proteomics 2021;232(1):104023.
[409] Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine 2020;61 (1):103030.
[410] Yeh YL, Su MW, Chiang BL, Yang YH, Tsai CH, Lee YL. Genetic profiles of transcriptomic clusters of childhood asthma determine specific severe subtype. Clin Exp Allergy 2018;48(9):1164–72.
[411] Rolland DCM, Basrur V, Jeon YK, McNeil-Schwalm C, Fermin D, Conlon KP, et al. Functional proteogenomics reveals biomarkers and therapeutic targets in lymphomas. Proc Natl Acad Sci USA 2017;114(25):6581–6.
[412] Niu L, Thiele M, Geyer PE, Rasmussen DN, Webel HE, Santos A, et al. Noninvasive proteomic biomarkers for alcohol-related liver disease. Nat Med 2022;28(6):1277–87.
[413] Poon W, Kingston BR, Ouyang B, Ngo W, Chan WCW. A framework for designing delivery systems. Nat Nanotechnol 2020;15(10):819–29.
[414] Mitchell MJ, Billingsley MM, Haley RM, Wechsler ME, Peppas NA, Langer R. Engineering precision nanoparticles for drug delivery. Nat Rev Drug Discov 2021;20(2):101–24.
[415] Li J, Esteban-Fernández de Ávila B, Gao W, Zhang L, Wang J. Micro/nanorobots for biomedicine: delivery, surgery, sensing, and detoxification. Sci Robot 2017;2(4):eaam6431.
[416] Ong AT, Serruys PW. Technology insight: an overview of research in drug-eluting stents. Nat Clin Pract Cardiovasc Med 2005;2(12):647–58.
[417] Bhatia SN, Chen X, Dobrovolskaia MA, Lammers T. Cancer nanomedicine. Nat Rev Cancer 2022;22(10):550–6.
[418] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019;18(6):463–77.
[419] 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(5):435–41.
[420] Chen C, Yaari Z, Apfelbaum E, Grodzinski P, Shamay Y, Heller DA. Merging data curation and machine learning to improve nanomedicines. Adv Drug Deliv Rev 2022;183(1):114172.
[421] Reker D, Rybakova Y, Kirtane AR, Cao R, Yang JW, Navamajiti N, et al. Computationally guided high-throughput design of self-assembling drug nanoparticles. Nat Nanotechnol 2021;16(6):725–33.
[422] Shamay Y, Shah J, Işık M, Mizrachi A, Leibold J, Tschaharganeh DF, et al. Quantitative self-assembly prediction yields targeted nanomedicines. Nat Mater 2018;17(4):361–8.
[423] Lu Y, Aimetti AA, Langer R, Gu Z. Bioresponsive materials Nat Rev Mater 2016;2(1):16075.
[424] Santana R, Zuluaga R, Gañán P, Arrasate S, Onieva E, González-Díaz H. Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. Nanoscale 2020;12(25):13471–83.
[425] Owh C, Ow V, Lin Q, Wong JHM, Ho D, Loh XJ, et al. Bottom-up design of hydrogels for programmable drug release. Biomater Adv 2022;141 (1):213100.
[426] Boztepe C, Künkül A, Yüceer M. Application of artificial intelligence in modeling of the doxorubicin release behavior of pH and temperature responsive poly (NIPAAm-*co*-AAc)-PEG IPN hydrogel. J Drug Deliv Sci Technol 2020;57(1):101603.
[427] Stiepel RT, Pena ES, Ehrenzeller SA, Gallovic MD, Lifshits LM, Genito CJ, et al. A predictive mechanistic model of drug release from surface eroding polymeric nanoparticles. J Control Release 2022;351(1):883–95.
[428] Jayatunga MKP, Xie W, Ruder L, Schulze U, Meier C. AI in small-molecule drug discovery: a coming wave? Nat Rev Drug Discov 2022;21(3):175–6.
[429] Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet 2020;395(10223):e30–1.
[430] Kirkpatrick P. Artificial intelligence makes a splash in small-molecule drug discovery. News Feature 2022.
[431] Zhang C, Mou M, Zhou Y, Zhang W, Lian X, Shi S, et al. Biological activities of drug inactive ingredients. Brief Bioinform 2022;23(5):bbac160.
[432] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36(5):421–7.