Fundamental Research 5 (2025) 1273–1287

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Review

Role of artificial intelligence in revolutionizing drug discovery

Ashfaq Ur Rehman[a,b,1], Mingyu Li[a,1], Binjian Wu[a,1], Yasir Ali[c,1], Salman Rasheed[d], Sana Shaheen[e], Xinyi Liu[a,f], Ray Luo[b], Jian Zhang[a,f,*]

[a] Medicinal Chemistry and Bioinformatics Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
[b] Departments of Molecular Biology and Biochemistry, University of California Irvine, Irvine, CA 92697, United States
[c] Institute of Chemistry, Slovak Academy of Sciences, 845 38 Bratislava, Slovakia
[d] National Center for Bioinformatics, Quaid-e-Azam University, Islamabad 44000, Pakistan
[e] Key Department of Biochemistry, Abdul Wali Khan University Mardan, Khyber-Pakhtunkhwa 23200, Pakistan
[f] State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

ARTICLE INFO

Article history:
Received 30 December 2023
Received in revised form 9 April 2024
Accepted 24 April 2024
Available online 9 May 2024

Keywords:
Artificial intelligence
Machine learning
Disease identification
Drug design
Drug discovery

ABSTRACT

The application of artificial intelligence (AI) in medicine, particularly through machine learning (ML), marked a significant progression in drug discovery. AI acts as a powerful catalyst in narrowing the gap between disease understanding and the identification of potential therapeutic agents. This review provides an inclusive summary of the latest advancements in AI and its application in drug discovery. We examine the various stages of the drug discovery process, starting from disease identification and encompassing diagnosis, target identification, screening, and lead discovery. AI's capability to analyze extensive datasets and discern patterns is essential in these stages, enhancing predictions and efficiencies in disease identification, drug discovery, and clinical trial management. The role of AI in expediting drug development is emphasized, highlighting its potential to analyze vast data volumes, thus reducing the time and costs associated with new drug market introduction. The importance of data quality, algorithm training, and ethical considerations, especially in patient data handling during clinical trials, is addressed. By considering these factors, AI promises to transform drug development, offering significant benefits to patients and society.

1. Introduction

The traditional drug discovery process is a complex and challenging endeavor that can require up to 15 years and over $1 to 2 billion for each approved drug [1]. This is primarily due to rising attrition rates and extended clinical trial duration [2]. Despite considerable investment in resources, almost 90% of potential drug candidates fail even after they have advanced to the phase-I clinical trial [3]. Advancing a drug candidate to phase-I clinical trial after rigorous optimization at the preclinical stage is considered a significant milestone for both pharmaceutical companies and academic institutions [4].

Large-scale computational screening and docking have been employed to enhance the success rate of lead compounds in clinical trials [5]. However, these methods have limitations such as inefficiency and inaccuracy [6]. To overcome these challenges, deep learning (DL) and ML algorithms, which are subsets of AI, have been identified as potential solutions [7]. These AI tools possess the ability to predict macrosystem properties with high accuracy while incurring low computational costs. As a result, there has been an increasing adoption of AI algorithms in the drug discovery process by chemical and biological scientists.

ML is extensively used in drug discovery, employing algorithms such as DL, Bayesian network (BN), random forest (RF), clustering, and support vector machine (SVM). The broad categorization of ML can be seen in Fig. 1. DL models process and analyze large amounts of data in tasks such as clinical imaging [8,9], virtual screening (VS) [10,11], and bioactivity predictions [12]. BNs predict toxicity or bioactivity and patients' response to treatments [13]. RF models are used for molecular target identification and feature selection, while clustering identifies patterns or relationships within data [14]. SVM is a supervised learning algorithm used to classify data into categories, with applications such as predicting pharmacokinetic properties, VS, and toxicity prediction [15].

Computational modelling based on AI and ML has made various drug discovery processes achievable, including chemical compound identification, target identification, peptide synthesis, drug toxicity and physiochemical property assessment, drug monitoring, drug efficacy and effectiveness assessment, bioactive agent prediction, protein-protein

* Corresponding author.
E-mail address: Jian.zhang@sjtu.edu.cn (J. Zhang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fmre.2024.04.021

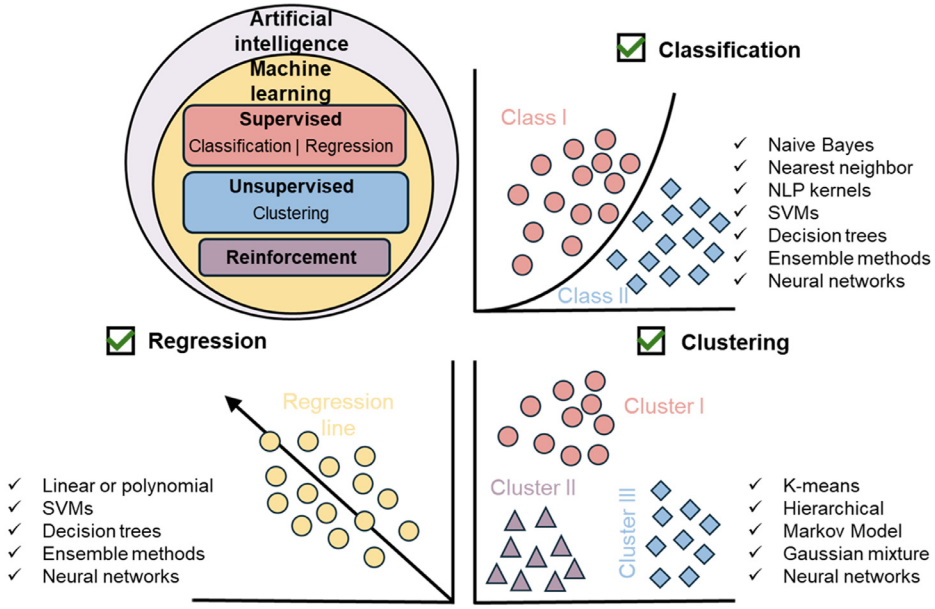

Fig. 1. Artificial intelligence and subtypes. Machine learning can be broadly categorized into supervised, unsupervised, and reinforcement learning, which can be further divided into several subcategories. For supervised learning, these include classification, where algorithms categorize input data into predefined classes, and regression, aimed at predicting continuous outcomes from input features. Additionally, unsupervised techniques such as clustering are used to grouping similar data points. These subcategories are fundamental to the architecture of machine learning algorithms, significantly contributing to diverse applications in drug discovery.

interaction, protein folding and misfolding identification, structure and ligand-based virtual screening (LBVS), QSAR modeling, and drug repositioning [16]. ML algorithms have made it faster to identify lead compounds from chemical libraries that contain over 106 million compounds and solve the toxicity challenges caused by off-target interactions [17]. Furthermore, ML-based tools such as AlphaFold have made it easier to predict the 3D structure of a target protein, which is crucial in drug discovery process.

AlphaFold is a recently developed AI-based tool from Google's DeepMind. Researchers have also used AI to discover novel peptides for therapeutic purposes. Provenzano et al. [18] created Deep-AmPEP30, a DL-based platform for the identification of short anti-microbial peptides (AMPs) [18], while Yi et al. [19] devised ACP-DL, a DL-based tool using the LSTM algorithm, for the discovery of novel anti-cancer peptides (https://github.com/haichengyi/ACPDL) [20]. AI is increasingly used in determining the proper drug dosage. Shen, Liu et al. [21] created the AI-PRS platform, a neural network-driven methodology that uses a parabolic response curve (PRS) to associate drug combinations and dose to efficacy [22], while machine and statistical learning techniques including k-nearest neighbor (kNN), Naive Bayesian, SVM, ANN, decision trees (DT), and RF are employed to forecast the hindrance in protein-protein interactions [23]. VS is an efficient method in computer-aided drug design (CADD), which involves screening a promising therapeutic compound from a pool of compounds.

ML can be used for VS with a filtered dataset, employing algorithms such as SVM, RF, and DT [24,25]. After validating the trained model for accuracy, it is used on new data sets to screen for compounds that have the desired activity against a target. The shortlisted compounds undergo ADMET (absorption, distribution, metabolism, excretion and toxicity) analysis and various bioassays before entering clinical trials. ML has the power to speed up VS, make it more robust, and even reduce false positives in VS. PARASHIFT, HEX, USR, and Shape algorithms have been constructed for LBVS. In recent years, with the rise of AI algorithms in the healthcare and pharma industry, different tools and models have been developed for both LBVS and structure-based virtual screening (SBVS), i.e., for LBVS, tools include SwissSimilarity, METADOCK, and HybridSim-virtual screening, while for SBVS, tools include Gypsum-DL, ENRI, and SPOT-ligand 2. Drug repositioning involves investigating drugs developed for one diseased condition and repositioning them for other diseased conditions in drug designing and discovery. This approach may be successful due to the possibility of multiple-target involvement in multiple diseases. The emergence of AI-based tools and algorithms in drug discovery provides a platform for future research, and different AI-based algorithms and web-based tools have been developed in recent years, such as DRIMC, DrugNet, DPDR-CPI, PHARMGKB, PROMISCUOUS 2.0, and DRRS.

This review aims to provide a thorough overview of AI's role in drug discovery (Fig. 2). AI can aid in various stages of drug discovery in various ways, including disease identification, target acquisition, computational screening, predicting drug toxicity, gene editing for developing gene therapies, and AI-based modeling for personalized drug dosing. We examine the current state-of-the-art AI technology and its potential to revolutionize drug discovery.

2. AI-based disease identification

AI has shown great potential in identification of infectious diseases. By analyzing large amounts of data from various sources such as EHRs (electronic health records), social media, and news reports, AI can quickly detect outbreaks of infectious diseases and provide early warning systems. AI can also assist in predicting the spread of diseases by identifying populations at high risk and tracking the movement of infected individuals. Due to its capacity for swift and precise data processing of enormous amounts of data, AI can significantly improve our ability to identify and respond to infectious diseases.

In recent years, the field of AI has shown remarkable progress in disease diagnosis, revolutionizing the way healthcare is delivered. AI technologies such as ML and DL have enabled medical professionals to develop highly accurate and reliable diagnostic models for various diseases. Early detection, precise diagnosis, and individualized treatment plans have all been made possible by the use of AI in disease diagnosis, which has improved patient outcomes. Infectious and non-communicable disease diagnosis using AI has recently made significant strides.

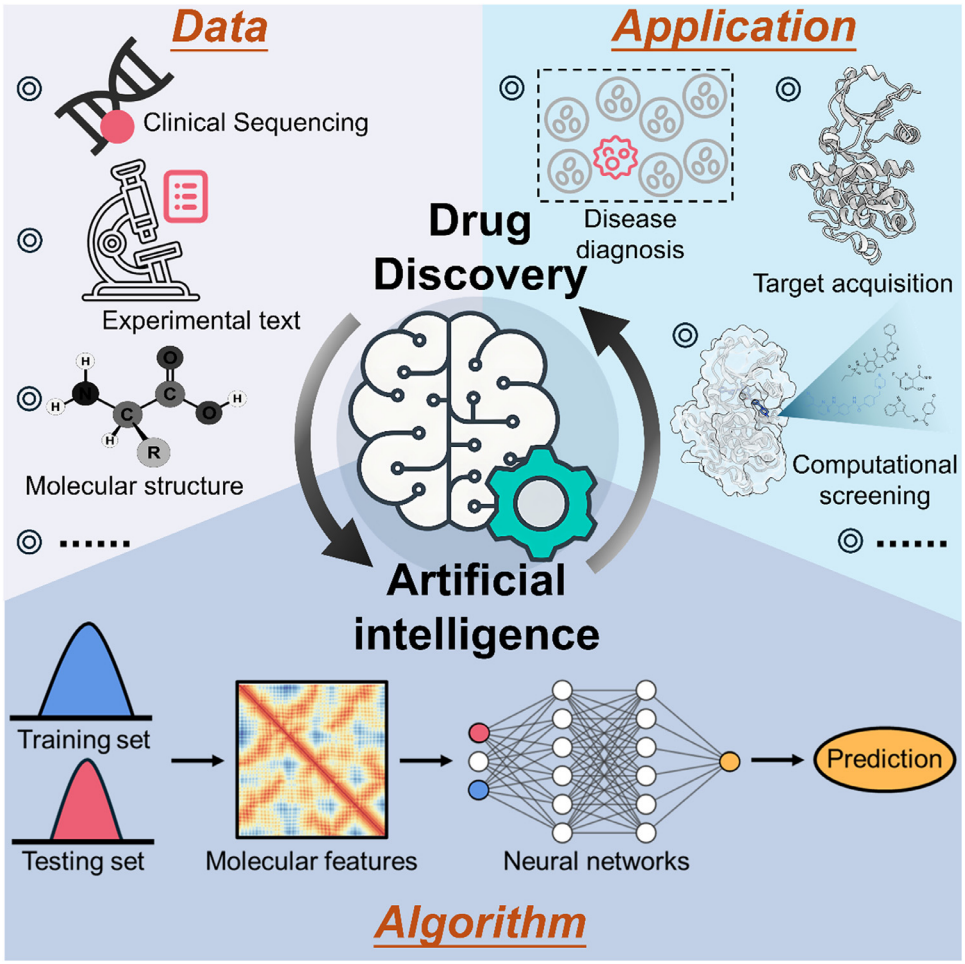

Fig. 2. Schematic representation of the integration of artificial intelligence (AI) in the drug discovery process. The diagram illustrates the workflow from data collection, encompassing clinical sequencing, experimental text, and molecular structure analysis, to the implementation of AI algorithms and neural networks. The applications of this process include disease diagnosis, target acquisition, and computational screening. This illustration underscores the transformative potential of AI in advancing medical research and healthcare solutions.

AI methods have greatly advanced the diagnosis of infectious diseases [26]. An exemplary case is the clinical decision support system (CDSS) known as "Sepsis Watch" to detect Sepsis early [27], a severe medical condition that occurs when an infection in the body triggers a chain reaction, resulting in a life-threatening medical emergency, often originating from infections in the lungs, urinary tract, skin, or gastrointestinal tract [28]. Sepsis Watch utilizes a unique ML method that combines multitask Gaussian processes (MGPs) with recurrent neural networks (RNNs) to identify sepsis [28]. The MGP component of the system learns the distributions of continuous functions for each dynamic variable. The system works by sampling dynamic features from the MGP hourly, which are then combined with static features and input into the RNN, a form of DL particularly adept at processing time-series data and essential for incorporating both static and dynamic features of hospital patient encounters [28]. This process generates a sepsis risk score between 0 and 1 for each patient. Additionally, the system includes scripts optimized to run every 5 min to identify patients who meet the sepsis criteria, facilitating early detection and prompt medical intervention [28]. Sepsis Watch has been trained using 50,000 patient records with both static (eg. prehospital patient comorbidities) and dynamic (eg. medication administrations) input features. It has been shown to improve the care of sepsis patients [28]. However, this study had a few limitations, including some false positive predictions that prompted clinical action even though the patient did not ultimately develop sepsis. Additionally, the study was limited to emergency department cases.

During the COVID-19 pandemic, there was a particular focus on the development of AI models for the effective diagnosis of this disease [29,30]. Chest X-ray is one of the efficient indirect methods for COVID-19 diagnosis. It was used to diagnose pneumonia associated with this viral infection [31]. Many ML models have been developed to predict the presence or absence of patterns in X-ray radiographs. For example, Narin et al. proposed an automated convolutional neural network (CNN) based diagnostic model for detecting pneumonia caused by coronavirus [9]. They developed pre-trained AI models using the X-ray radiographs of healthy individuals, patients with COVID-19, patients with viral pneumonia, and patients with bacterial pneumonia. The reported accuracies in classification reached up to 96%. Aapka Chikitsak is an AI-powered conversational bot designed to enhance telehealth services in India by providing accessible COVID-19 information and addressing the imbalance between the demand for and the supply of human healthcare providers. Initially, the user's query is converted from audio input into text, and the text is used as a basis for performing Natural Language Understanding to decode the semantic meaning. Subsequently, relevant entities are identified and linked to their corresponding intents in Dialogflow. The bot then generates a response, converts it to speech, and delivers it back to the user [32].

ML models have also been built to assist in the diagnosis of other infectious diseases such as urinary tract infections (UTIs), which are often associated with high rates of diagnostic errors. Taylor et al. reported a retrospective cohort analysis of approximately 80,387 adults who vis-

ited the emergency department with UTI symptoms [33]. Considering symptoms as well as blood and urine sample analyses, six AI algorithms were developed for the diagnosis of UTI: SVM, artificial neural network (ANN), elastic net, adaptive boosting (AdaBoost), RF, and extreme gradient boosting (XGBoost). The models were built using a full set of 211 factors and a reduced set of 10 variables, e.g., gender, epithelial cells in the urine, history of UTI, and age. The XGBoost showed superior accuracy over the other algorithms, with an area under the curve (AUC) of 0.88 and 0.90 for the full and reduced XGBoost models, respectively. The sensitivity and specificity were 61.70% and 94.90% for the full, and 54.70% and 94.70% for the reduced models, respectively [33]. Lifestyle disorders, such as diabetes, obesity, and hypertension, are associated with the way people live, their diet, exercise levels, etc.

ML models have achieved substantial success in the prevention and detection of HIV, a virus causing human progressive immune system failure and promoting cancers. Xiang et al. developed an ensemble approach combining GCN (graph convolutional network) with LR and RF, aiming to identify individuals at high risk of HIV infection for prevention [34]. They built a social network in which each node stood for an individual participant, and the edges between them represented social connections. Each node was assigned with a feature set, including sociodemographic characteristics and sexual behavioral characteristics. The ensemble methods produced promising results on HIV infection detection (GCN+LR with accuracy of 93.4% and F1 of 88.4%; GCN+RF with accuracy of 96.6% and F1 of 94.6%). Heerden et al. created a conversational agent guiding users through an HIV counseling and testing session utilizing NLP (nature language process). This agent encourages the targeted population to openly discuss their concerns with a virtual assistant, and its effectiveness has been confirmed through human evaluation [35]. Compared with previous agents, this agent made testers feel like the session was more private and anonymous with more gentle language and more accurate detection.

Many AI-based algorithms have been developed for the early prediction and management of diabetes. For instance, Spänig et al. developed an interactive AI model with the capability of speech recognition and speech synthesis that acts as a virtual doctor, interacts directly with patients. An open-source system CMUSphinx is utilized to develop robust speech recognition capabilities. To support localized speech recognition, the essential German language data is obtained, including a German language model, an acoustic model, and a dictionary from the VoxForge dataset [36], which aggregates transcribed speech specifically for use in speech recognition technologies. This virtual doctor predicts Type-2 diabetes mellitus with an AUC of 0.84 [37]. The incidence of retinopathy is high among diabetic patients. Gulshan et al. developed a deep CNN model that bypasses the human capacity at interpreting, evaluating, and classifying retinal images. The model is trained using 128,175 retinal photographs which are evaluated by a panel of clinicians and ophthalmologists. The model is demonstrated to have a high sensitivity and specificity of 97.50% and 93.40%, respectively [38]. In 2018, the U.S. Food and Drug Administration approved the marketing of the first AI-based medical device called IDx-DR [39] for detecting diabetic retinopathy. The device has a retinal camera through which the retinal image of the patient is taken and analyzed. The device is autonomous and decides on one of the following results based on the image quality (i) “more than mild diabetic retinopathy detected: refer to an eye care professional” or (ii) “negative for more than mild diabetic retinopathy; rescreen in 12 months” [39].

Alzheimer’s disease is a neurodegenerative disorder in the brain. Genetic factors and age are major risk factors associated with AD. However, recent research indicates that other factors, such as environmental and lifestyle factors, can also contribute to the development of AD [40]. AI-based algorithms have shown promising results in the early prediction and diagnosis of AD [41]. For instance, Jo et al. proposed a hybrid model that combines DL -based feature extraction with ML algorithms for AD diagnosis using magnetic resonance imaging (MRI) scans [42]. The proposed model achieved a classification accuracy of 96%. Similarly, Shi et al. developed an AI-based model for the early prediction of AD using positron emission tomography (PET) images [43]. The model achieved an accuracy of 89.5% in predicting the onset of AD within two years. Cancer is another NCD that has benefited from the recent advancements in AI-based algorithms. AI-based algorithms have been developed for various cancer-related tasks, such as diagnosis, prognosis, and treatment planning. For example, Ding et al. developed a DL algorithm that interprets PET of the brain for the early prediction of AD. Their model showed a specificity and sensitivity of 82% and 100%, respectively. This model can predict AD, on average, 75.8 months before its diagnosis, with a ROC (receiver operating characteristic) of 0.98 [8]. Li et al. developed a DL-based model that automatically evaluates dementia severity by analyzing resting-state functional magnetic resonance imaging data (rs-fMRI). The study involved 133 patients with Alzheimer’s disease, and their clinical dementia rating (CDR) scores ranged from 0.5 to 3. To extract features, three-dimensional CNNs were applied to rs-fMRI data. The accuracy of the model was found to be highly satisfactory [44]. In addition, AI-based algorithms have also shown promising results in predicting the response to treatment in cancer patients. For example, AI has the potential to improve the speed of analysis and the accuracy of image interpretations. Esteva and co-workers developed a CNN model trained with images of skin lesions to classify different types of skin cancer [45]. Albayrak et al. used deep learning to extract features from breast histopathological images to detect mitosis. The proposed model extracted CNN features for SVM training and detected breast mitosis [46]. Causey et al. developed CNN model-based algorithm, NoduleX to predict malignant lung nodules from clinical CT data. The model was trained using over 1000 lung nodule images from LIDC and IDRI. NoduleX predicted with a 0.99 AUC [47]. Shiri et al. tested ML methods using Radiomics analysis for predicting EGFR and KRAS mutation status in NSCLC (non-small cell lung cancer) patients that showed AUCs of 0.82 and 0.83, respectively [48]. CNN models also power histological image analysis to diagnose cancer [49-51].

AI algorithms have demonstrated promising outcomes in the diagnosis, prediction, and management of non-communicable diseases including diabetes, Alzheimer’s disease, and cancer. The implementation of AI-based algorithms can facilitate the early detection of these diseases, thus enabling timely interventions and personalized treatment plans, ultimately leading to improved patient outcomes. By leveraging the power of AI, healthcare professionals can potentially reduce the burden of NCDs, enhance the quality of healthcare delivery, and optimize healthcare resources. The detailed overview of role of AI in disease identification can be seen in Fig. 3.

3. Target identification

Target identification is a critical step in the drug discovery process. The traditional approach involves time-consuming and costly experimental methods, such as high-throughput screening (HTS) and X-ray crystallography. However, the use of AI has revolutionized this field by enabling the identification of potential targets through computational methods. Fig. 4 illustrates how AI-aided role in drug discovery.

AI-based target identification involves the use of ML algorithms to analyze large datasets and identify targets with the potential to interact with a given drug. This approach utilizes various data sources, such as gene expression profiles, protein-protein interaction networks, and biological pathways, to generate a list of candidate targets [52]. ML algorithms, such as SVMs and neural networks, can then be used to prioritize these targets based on their relevance to the disease of interest.

Furthermore, AI-based target identification can help identify novel targets that were previously unknown or overlooked. By analyzing large datasets from various sources, ML algorithms can uncover hidden patterns and relationships that may not be immediately apparent using traditional methods. This can lead to the discovery of new biological pathways and targets that may have therapeutic potential. AI-based target identification has the potential to revolutionize the drug discovery

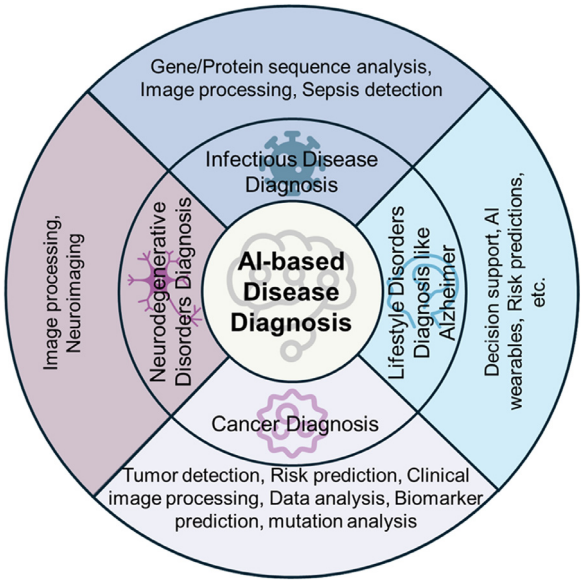

Fig. 3. The overview of AI-based disease identification. AI-based disease identification can be divided into four categories: neurodegenerative disorders diagnosis, cancer diagnosis, infectious disease diagnosis and lifestyle disorders diagnosis.

process by enabling the identification of potential targets through computational methods. There are multiple experimental methods available for identifying drug targets, such as affinity pull-downs and genome-wide knockdown screens. However, these approaches require a significant amount of labor, resources, and time, and are also subject to a high rate of failure. In contrast, computational methods have the potential to significantly decrease the effort and resources required for drug target identification [53].

Cryo-EM, or electron cryo-microscopy, revolutionized the investigation of proteins and protein complexes in the 1980s. Single-particle analysis (SPA), electron cryo-tomography (cryo-ET), microcrystal electron diffraction (MicroED), low-energy electron holography (LEEH), and cryo-scanning transmission electron tomography (CSTET) are a few examples of cryo-EM imaging techniques [54,55]. For a deeper understanding of the structures of protein complexes in their environment, cryo-electron tomography (cryo-ET) has shown great promise. Computer algorithms are used to combine these images to create a 3D structure representation.

The software CryoDRGN, created by Zhong et al. [56], uses ML to enable the reconstruction of proteins and protein complexes from heterogeneous cryo-electron microscopy data [56]. In order to embed heterogeneous single-particle cryo-EM images in a low-dimensional latent space and produce 3D volumes as a function of this embedding, the authors have proposed a technique that makes use of ML models. CryoDRGN can produce an infinite number of maps from the imaged ensemble and is capable of modeling complex ensembles with both continuous and discrete heterogeneity. The software can also visualize the motion of the protein [57].

Trypsin is one of the most important and widely used proteolytic enzymes in mass spectrometry (MS)-based proteomic research. The digestion of proteins by protease enzyme is a basic step in the protein identification using MS. A few AI tools were developed to efficiently predict the digestion behavior of the protease enzymes [58,59]. DeepDigest is the first algorithm developed using a DL method to predict the proteolytic cleavage sites of eight different protease enzymes (trypsin, ArgC, chymotrypsin, GluC, LysC, AspN, LysN, and LysargiNase). The DL model was trained on 19 public large-scale data sets covering the eight proteases from samples of four organisms (E. coli, yeast, mouse, and human). The predictive ability of the tool was evaluated by the AUC, F1 scores, and the Matthews correlation coefficients (MCCs); the values were 0.956–0.98, 0.66–0.90, and 0.65–0.84, respectively [58]. Sun et al. developed an algorithm to predict the missed cleavage site in the tryptic protein digestion [59]. The algorithm is demonstrated to have a high accuracy, with an AUC of 0.99. This algorithm can be incorporated into the peptide database search in the MS analysis to facilitate the identification of proteins more effectively and efficiently.

The development of computational tools, high-performance computers, and ML algorithms is not limited to three dimensional (3D) models of protein targets but also enabled a large number of drug discovery tools. This is a significant advancement in experimental techniques that are fraught with challenges. For example, the X-ray diffraction technique is limited to crystallizable samples, which is a major experimental limitation [60].

In the absence of experimental data, computational techniques have been used for decades to forecast the three-dimensional structure of

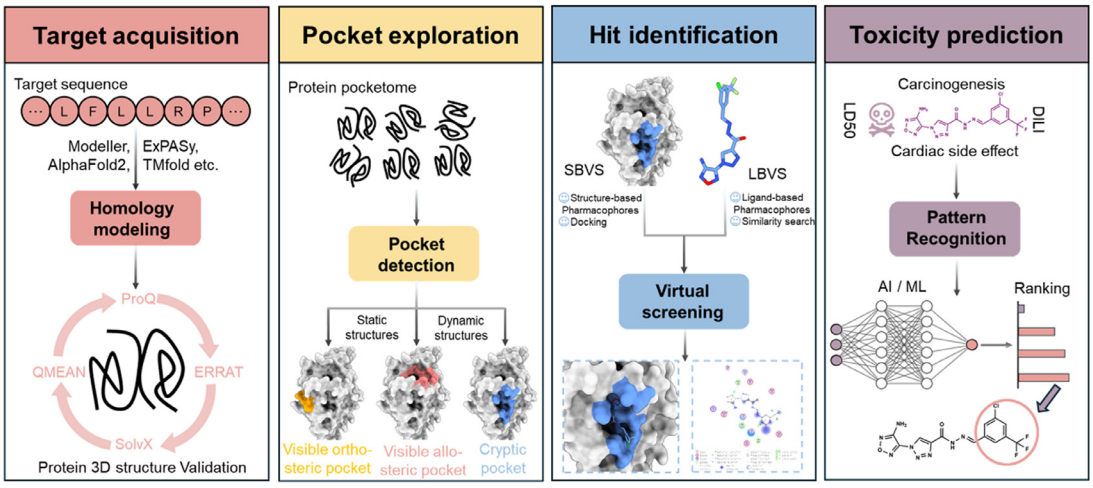

Fig. 4. The overview of the AI-based process for identifying and evaluating small molecules. Beginning with target acquisition, where protein sequences are modeled using tools like Modeller and AlphaFold2, followed by Homology Modeling and validation of protein 3D structure through ProQ and SolvX. Next is the pocket exploration, including the identification of both orthosteric and allosteric binding sites within the protein proteome, via static and dynamic structures. The hit identification stage employs SBVS and LBVS to pinpoint potential compounds. Finally, the toxicity prediction stage uses pattern recognition techniques in AI/ML to predict the ADMET properties.

proteins. Modeller, a homology modeling software, predicts a protein's structure based on its alignment to one or more proteins of known structure (templates) [61]. AlphaFold, the DL algorithm developed by DeepMind, a UK based company, predict the 3D structure of proteins from their amino acid sequences [62]. The development of this AI- tool is claimed to be a breakthrough in drug discovery as it was used to solve the structure of nearly 200 million proteins, which is ~98.5% of the proteins in the human body. In July 2021, the predicted three-dimensional models for the whole human proteome generated using AlphaFold, were made available to the public, as recently reported in Nature [63]. Together with the European Bioinformatics Institute (EMBL-EBI), a database called AlphaFold DB (https://alphafold.ebi.ac.uk/) was created to store all the structures solved so far with AlphaFold. However, the effect of mutation on the folding of proteins is beyond the capability of AlphaFold [64]. It is also limited to predicting only a single state of a given protein, it does not consider the dynamic nature of protein structures [65]. Recently, new methods based on natural language processing (NLP) which only uses the amino acid sequences from sequence databases to learn structural, functional and evolutionary patterns and predict structural conformation. In 2022, two methods gained attention i.e. ESMFold and EMBER3D. ESMfold, developed by Lin et al. in 2022, utilizes a masked transformer protein language model with a deep understanding of biological properties, trained with 15 billion parameters. While it falls short of AlphaFold2 in overall performance, as indicated by lower TM-scores (0.68 compared to AlphaFold2's 0.85 on CASP14), ESMFold outperforms AlphaFold2 when evaluating the amino acid sequence alone without multiple sequence alignment (MSA) (0.68 versus 0.37 on CASP14). ESMFold demonstrates comparable accuracy to AlphaFold2 for structures predicted with high confidence, exhibiting a median all-atom RMSD (root-mean-square deviation) of 1.91Å and a backbone RMSD of 1.33Å—achieving accuracy levels akin to experimental results. Additionally, ESMFold shows a substantial improvement in prediction speed, eliminating the need for MSA construction. The authors leverage this approach to introduce the ESM Metagenomic Atlas, predicting over 617 million structures from metagenomic databases. Among these, 225 million structures are predicted with high confidence, including novel ones [63]. While Ember3D falls short of surpassing AlphaFold in performance, it exhibits significantly faster processing speeds compared to both AlphaFold and ESMFold. Notably, AlphaFold2 struggles to efficiently investigate the impact of single amino acid variants on protein structure. In contrast, Weissenow et al. [64] demonstrated that Ember3D's predicted distance maps show a strong correlation with native and mutant 3D structures obtained through deep mutational scanning, outperforming ESMFold in this regard. The researchers also developed a tool that highlights differences between native and mutant predicted structures for all possible amino acid exchanges at each position in a protein sequence. The tool utilizes the similarity between the native and mutated amino acids to identify exchanges that may have a significant impact on the protein structure [64]. Cheng et al. introduced AlphaMissense, a DL model that extends the capabilities of the AlphaFold2 protein structure prediction tool. This model undergoes training with population frequency data and incorporates both sequence information and predicted structural context to enhance its overall performance. AlphaMissense effectively categorizes 32% of all missense variants as likely pathogenic and identifies 57% as likely benign, achieving a 90% precision cutoff on the ClinVar dataset. This outcome ensures a robust prediction for the majority of human missense variants [65]. A specific and detailed review [66] on all the recent advances in protein structure prediction are reviewed by Peng et al. Some breakthrough protein structure prediction models are tabulated in Table 1.

In addition, AI algorithms can predict the likelihood of allosteric regulation in a protein based on its structure and ligand presence, and identify allosteric modulators that may modify its activity [66]. ML methods, like Allosite; developed by our group [67], and AlloPred [68], used SVM with optimized features for protein pocket classification [69], while others used RF [70] to build a three-way predictive model. Advanced ML models, such as XGBoost [71], can now classify allosteric sites with greater accuracy. XGBoost implements the gradient boosting algorithm with regularized terms to reduce overfitting and has demonstrated superior predictive performance in protein-protein interactions [72] and hot spots [73] compared to SVM and RF.

Tian et al. developed a webserver called PASSer (Prediction of Allosteric Sites Server), combining the results of XGBoost and GCN to predict the allosteric sites on proteins using an ensemble learning method. A total of 1946 entries information of allosteric sites from Allosteric Database were collected for training with 19 descriptors extracted for each site by Fpocket. For a given pocket, physical properties are calculated and fed into the XGBoost model while an atomic graph is fed into the GCNN model. The final result is the averaged probability of these two models. The model showed an accuracy of 0.97, precision of 0.73, and specificity of 0.98. The aforementioned model can acquire knowledge about both the physical traits and topology of allosteric pockets and has been demonstrated to outperform the XGBoost and GCN models individually. The findings are consistent with earlier research and have a greater likelihood of placing allosteric sites in top positions. The online server is equipped with an easily navigable interface. Protein structures and top pockets are displayed in an interactive window on the result page [66].

Protein structural fluctuations generate novel cryptic pockets [74-77], which offer druggable sites beyond experimentally determined structures. Cryptic pockets regulate protein functions allosterically. These pockets are hidden protein cavities that open up when ligands or protein partners bind [78]. Targeting these cryptic pockets offers drug development opportunities. Cryptic pockets can target undruggable proteins [79]. An algorithm to predict which proteins have cryptic pockets could help prioritize proteins to target in cases where proteins lack ground state pockets or modulators are difficult to design. CryptoSite is an excellent supervised ML algorithm that predicts ligand-binding cryptic pockets from protein structures [80]. CryptoSite accurately predicts cryptic pocket participation for amino acid residues (ROC-AUC = 0.83). Miller et al. developed PocketMiner, a graph neural network (GNN), to predict cryptic pockets in protein structures. The model is trained using residues likely to form cryptic pockets from 2400 simulations of 35 proteins. Model AUC was 0.87. This supports molecular dynamics simulations for cryptic pocket identification [78].

Madhukar et al. developed BANDIT, a Bayesian ML platform for drug target prediction that integrates multiple data types to achieve greater predictive power. BANDIT utilizes over 20 million data points from six distinct data types, including drug efficacies, post-treatment responses, bioassay results [81,82], and known targets [83,84]. The platform achieves high accuracy at identifying shared target interactions and uncovers novel targets for cancer treatment. BANDIT was tested on ~2000 compounds and can quickly pinpoint potential therapeutics with novel mechanisms of action to accelerate drug development.

In 2021, Kozlovskii and Popov developed a DL approach to predict the binding site for small molecules on nucleic acids, DNA, and RNA, based on their 3D structures [85]. Their approach is called BiteNetN (https://sites.skoltech.ru/imolecule/tools/bitenet/) having a large dataset of 1933 nucleic acid-ligand complexes, including 1065 DNA and 886 RNA structures (18 structures contain both DNA and RNA) of different type. It was the first 3D CNN to learn features directly from nucleic acid structures. They validated the model using two different protein structures, HIV-1 transactivation response element RNA and ATP-aptamer structures. The model showed an AUCROC of ca. 0.87, proved to be top-performing for protein-small molecule and protein-peptide binding site detection [86,87].

DeepDTnet is a network-based DL method developed by Zeng et al. to aid the target identification process [88]. The model is trained with chemical, genomic, and cellular network data for the accurate prediction of molecular targets. The model is shown to have a high accuracy in prediction with an AUC of 0.96 [88]. In another study, Mamoshina et al. developed ML techniques to analyze human muscle transcriptomic

Table 1
List of the methods for protein structure prediction with their standlone availability links.

S. No	Method	Web links
1	AlphaFold2	https://github.com/deepmind/alphafold
2	ColabFold	https://github.com/sokrypton/ColabFold
3	DMPFold2	https://github.com/psipred/DMPfold2
4	ESMFold	https://github.com/facebookresearch/esm
5	EMBER3D	https://github.com/kWeissenow/EMBER3D
6	HelixFold-Single	https://paddlehelix.baidu.com/app/drug/protein-single/forecast
7	Openfold	https://github.com/aqlaboratory/openfold
8	OmegaFold	https://github.com/HeliXonProtein/OmegaFold
9	RoseTTAFold	https://github.com/RosettaCommons/RoseTTAFold
10	RNG	https://github.com/aqlaboratory/rgn
11	RNG2	https://github.com/aqlaboratory/rgn2

data to discover biomarkers associated with muscle-related diseases and to identify tissue-specific drug targets [89]. The authors collected transcriptomic data from human muscle tissues and used a combination of unsupervised and supervised ML algorithms to identify differentially expressed genes and gene modules that are associated with muscular dystrophy and sarcopenia. They further investigated the identified gene modules and pathways using functional annotation and network analysis tools to identify potential drug targets. Their best model showed a Pearson correlation of 0.80.

After identifying the protein targets, corresponding novel lead compounds can be further generated from scratch, leveraging the pocket features and ligand topology. Zhang et al. developed a pocket-aware ligand generator ResGen. A two-level autoregression protocol for molecular generation is introduced in the model to better incorporate the geometry of protein pockets. The global autoregression is to generate atoms in pockets, and the atomic autoregression is to produce the coordinates and topology of the newly added atoms. ResGen could generate more physically sensible molecules with tighter binding [90]. Wei et al. developed a fragment-based generation model Lingo3DMol built on the transformer-based structure [91]. A new molecule representation FSMILES is introduced, enabling the generation of 3D molecules with reasonable conformations and topology. Additionally, non-covalent interactions and ligand–protein binding patterns are also considered during the generation. Lingo3DMol demonstrates excellent performance in terms of drug likeness, synthetic accessibility, pocket binding mode and molecule generation speed.

4. AI-enabled virtual screening in drug discovery: opportunities and challenges

The initial phase of drug discovery usually involves computational screening of numerous compounds to identify those with the desired cellular or biochemical effects. To enhance the speed, efficiency, and cost-effectiveness of this process, new methods are constantly being developed. A positive response during the first round of screening in a biochemical assay identifies primary “hit” compounds. Subsequently, additional screening is performed to assess whether the physicochemical and pharmacological properties of the hit compounds are suitable for developing a medicine. If they pass this filter, they are designated as “leads”. These leads are then refined chemically and subjected to biological screening in subsequent rounds before proceeding to clinical testing. With some luck, a lead may ultimately receive drug approval, a process that may take 12–15 years from the beginning of testing [92].

Despite significant advancements in drug discovery and medicinal chemistry technologies, drug development still remains a slow and expensive process. The current standard process involves HTS, where *in vitro* assays are conducted on thousands of compounds to identify hit compounds that can be optimized to lead compounds with desirable properties such as increased potency, solubility, and reduced toxicity and off-target effects [93].

The conventional drug discovery process involves synthesizing and testing thousands of compounds, which is both time-consuming and costly, requiring large amounts of protein supplies and established laboratory methods for bioactivity testing. In contrast, VS has emerged as a cost-effective approach to scan millions of commercially available compounds and prioritize those for further testing, synthesis, or purchase. VS methods are classified into two categories: structure-based and ligand-based methods. However, despite the potential benefits, it still takes an average of 10 to 15 years and over $2 billion to develop a single drug [94].

4.1. Structure-based methods

To utilize structure-based methods, 3D structural information of the protein-ligand complex or at least the protein’s binding site is needed. Molecular docking is a commonly used technique that generates multiple possible binding poses of a ligand in the target protein structure and ranks them using a scoring function (SF) [95] to estimate their binding affinity [96]. Recently, machine/DL-based SFs [21] have been introduced as a new group of SFs. On the other hand, ligand-based methods such as QSAR modeling, molecular similarity search, and ligand-based pharmacophores, are more established technologies that require only ligand information [97], unlike structure-based methods. AI techniques can also be employed to improve the efficiency of computer-aided drug discovery processes, which often need extensive high-performance computing resources and significant computation time.

Gentile et al. reported an open-source protocol for AI-enabled VS methods to screen libraries with billions of molecules. They used a screening platform called Deep Docking (https://github.com/jamesgleave/DD_protocol) which can accelerate the SBVS by 100 folds. The input data consists of the molecule’s SMILES with its Morgan fingerprints as descriptors and the target’s structure. Deep Docking performs molecular docking for a small subset of a large library based on DNN to infer the ranking of the yet-unprocessed remainder, followed by ligand-based prediction of the docking for the rest of the library. In this way, Deep Docking discards undockable molecular structures without wasting computational resources. A key advantage of this protocol is that it can be used in conjunction with other docking programs such as Glide, Autodock-GPU, and FRED from OpenEye. Although, the deep docking method provides faster screening, it is limited to I) the availability of graphical processing units (GPU) and ii) the quality and accuracy of the docking program used [10].

Various ML techniques, including Naïve Bayesian (NB) classifiers, kNNs, SVMs, RFs, and ANNs, can be used for VS. While SVMs and ANNs are commonly regarded as the most accurate, each technique has its own strengths and weaknesses. For instance, NB excels at identifying favorable scaffold fragments, RFs can be parallelized and boosted, and kNN is simple to implement and can utilize MTL. Combining an ensemble of ML models is often preferred as it can enhance performance [98].

For LBVS, NB classifiers are widely utilized and exhibit excellent performance. Wang et al. investigated four classification models to identify

inhibitors of methicillin-resistant Staphylococcus aureus. These models include NB, SVM, recursive partitioning (RP), and kNN algorithms, incorporating various sets of physicochemical descriptors and fingerprints. Among these models, the NB classifier displayed the highest performance in the testing set of, achieving an accuracy of 0.891 and a specificity of 0.920 [99]. Likewise, Lian et al. [100] showed that NB models outperformed SVM models with an accuracy of 0.975 and a specificity of 0.989 in classifying the inhibitors and non-inhibitors of neuraminidase. They further identified nine effective inhibitors of neuraminidase using an enhanced ensemble model comprising NB models and SVMs.

Another simple ML classification method is kNN. Unlike NB, kNN is intuitive. In searching for G-Protein Coupled Receptor ligands, Luo et al. found that kNN-QSAR with variable selection outperformed LBVS approaches without ML [101]. Compared to other ML methods, kNN's performance is usually in the middle [99,102]. kNN is less popular than the ML methods below because of this. In screening for Estrogen Receptor-mediated endocrine disruptors, kNN has been used for VS [103].

Vapnik et al., introduced SVMs for the first time [104]. They function by representing input data as feature vectors and plotting them within the same-dimensional space. Although SVM can be used for unsupervised learning, supervised learning is preferred for VS because it guarantees that a compound will eventually be categorized as active or inactive. Chandra and colleagues conducted a study to identify potential inhibitors for PTP1B, a treatment for Type 2 diabetes [105]. Multiple ML models were developed, and the best model that utilized SVM was applied to an external database. This model successfully identified five inhibitory compounds, two of which demonstrated significant activity *in vitro*. In a separate study, Deshmukh et al., [106] discovered that their SVM model was able to identify almost half of the known FEN1 inhibitors in a test set, and also identified previously unknown inhibitors from the Maybridge small molecule database, which were experimentally validated. Baba et al. found that SVM models with regression were more effective than RF in predicting a compound's ability to permeate the skin[107]. Lee et al., [108] employed RF-QSAR to study compound polypharmacology, resulting in the creation of a targetfishing server that identifies possible targets for a given compound. Their method achieved an overall AUC score of 0.97 and outperformed NB-based methods in external testing [109].

4.2. Ligand-based virtual screening (LBVS)

LBVS is based on selecting, from databases, molecules that share similar structural features with an active ligand. Pharmacophore-based VS is one of the LBVS techniques. It involves building 2D fingerprints of one or more active ligands using molecular descriptors such as hydrogen-bond donors, hydrogen-bond acceptors, and aromatic rings. These 2D fingerprints are then used to identify molecules, from large chemical libraries, which have matching pharmacophoric features. ML also helps to study the correlation between molecular descriptors (or even atomic descriptors [110,111] and the biological activity of a ligand. This is a broad category of research known as Quantitative Structure-Activity Relationship (QSAR), where the activity of a ligand depends on its pharmacophoric features. Melge et al. developed hybrid inhibitors using the pharmacophore fingerprint of two well-known anti-cancer drugs Ponatinib and Vorinostat [112]. They developed a supervised ML approach for 2D-QSAR and 3D-pharmacophore studies to predict the inhibitory activity of novel hybrid molecules. The model had AUC values of 0.98 and 0.94 for the two different cancer targets, BCR-ABL and Histone deacetylase (HDAC), respectively. Based on in-vitro evaluations, the identified novel hybrid molecules showed the potential to develop into lead compounds. Dhamodharan et al. developed three AI models based on genetic function approximation (GFA), SVM, and ANN, to predict the activity of acetylcholinesterase (AChE) and Beta-Secretase 1 (BACE1) dual inhibitors for AD treatment [113]. The predictive power of the models was evaluated on a test set of 11 inhibitors of AChE and BACE1. The ANN model had the best predictive power with R^2 (coefficient of determination, a statistical measure within regression analysis) values of 0.85 and 0.78 for AChE and BACE1, respectively. Dhamodharan et al., [113] used a target-specific scoring model to identify potential inhibitors for 12 targets from the SAM MTase family. 446 actives and 1294 decoys were docked using Glide and the DUD-E website. The MLP outperformed other docking tools in a binary classification experiment.

Atomwise, Inc. developed AtomNet, one of the earliest CNNs utilized for VS [113]. Unlike most ML-based VS, AtomNet employs an SBVS and its architecture includes input and logistic-cost layers, four convolutional layers, and two fully connected layers. The filters in AtomNet's convolutional layers correspond to chemical functions, allowing the network to identify features that aid in binding. AtomNet consistently achieves AUC scores >0.74 on various benchmark datasets, outperforming numerous previous docking models. These features and capabilities, previously exclusive to NB and RF classifiers, now enable ANNs to identify features that aid in binding, making them more accurate and less of a black box in VS.

Alzheimer's disease is characterized by amyloid-beta (A) plaques and neurofibrillary tangles (NFT) of hyperphosphorylated tau protein. Acetylcholine (ACh) levels are lower in AD brains [114-116]. Tau may be phosphorylated by GSK-3, CDK5, and other Alzheimer's disease-related enzymes and targets [117-119]. Side effects from polypharmacy are possible. Drugs that target one protein frequently have side effects on other proteins as well. Through the inhibition of multiple targets in complex diseases like Alzheimer's, polypharmacology can increase a drug's effectiveness. With this technique, MTDL (Multi-Target Directed Ligands) screening was increased [43]. Using NB and RP classifiers in an LBVS on MTDLs, Fang et al. discovered compounds that bound to 25 targets, including BACE1, the M1 subtype of mAChR, APP, CDK5, and GSK-3 [120]. The model was validated using current AD medications, and it was used to forecast upcoming MTDLs. As scientists become more aware of its effectiveness, ML in VS for drug discovery will expand. The effectiveness and cost of drug discovery will be enhanced by computer science and medicinal chemistry.

ML-based screens are computationally efficient and successful in modern CADD, which requires vast computational resources to screen expanding chemical libraries due to automated synthesis and robotics. OpenEye GigaDocking and VirtualFlow are supercomputing platforms for docking large libraries that have screened billions of molecules using thousands of CPUs/GPUs in relatively short periods [121]. However, they are resource-intensive compared to Deep Docking (DD) [122]. DD requires fewer computational resources, making it an attractive alternative for large-scale VS.

Gentile et al. developed an open-source protocol for AI-enabled VS of billion-molecule libraries. Deep Docking (https://github.com/jamesgleave/DD_protocol), a screening platform that accelerates SBVS by 100-fold, was used. One of the fastest AI-enabled docking platforms and the only one tested on 1B+ libraries is DD. The DD protocol does not require a docking program, so it can be used with emerging large-scale docking methods to improve throughput. technical limitation of Deep Docking is GPU acceleration, which is needed for optimal performance unlike CPU-based docking platforms. The protocol only provides docking details for top-scoring molecules, ignoring large fractions of chemical libraries for fast VS. Docking campaigns assessing hit rate variability with docking scores [123] or rescoring low-ranked molecules [124] should consider a bruteforce approach. A docking program's ability to prioritize active molecules from an ultra-large library also determines the quality of DD results [123]. Bender et al.'s guidelines for large-scale docking benchmarking are helpful [122].

A wave of DL methods and applications has improved affinity—and other properties like ADMETox—prediction in VS over the last decade(s). Models learn characteristics rather than using human-designed descriptors. Novel encodings like voxels (where physicochemical atomic properties are pinned to locations in 3D space) and graphs (which describe bonded and non-bonded connectivity between atoms) appear to capture the variety of information needed for ligand-binding.

DeepAtom [125], a 3D grid-based method that assigns physicochemical properties to each grid cell, may be suitable for modeling protein-ligand binding complexity. The study [126] and DGraphDTA model [127] show that encoding chemical and biological objects as graphs works well. Despite this, several challenges remain, and new ones have emerged, including chemical encoding precision, generalization of chemical space, lack of (large and high-quality) data, model comparability, and interpretability.

Finding new therapeutic uses for already-approved drugs is a crucial aspect of drug repurposing. Due to their previously tested pharmacokinetics and toxicity, repurposed medications have a higher success rate [128,129]. Drug repurposing skips lead optimization and preclinical studies to enter phase-II trials. Due to complex disease pathophysiology, many drugs have off-target binding and are excluded from pre-clinical trials [130]. Reker and colleagues developed a method to predict molecular targets of known drugs and computer-generated de novo small molecules, including key-target and off-target proteins. Self-organizing map-based prediction of drug equivalence relationships (SPiDER) is this method. The software was trained on 12,661 manually curated active molecules [131]. The ROC ranged from 0.86 to 0.93 in a 10-fold cross-validation of SPiDER's predictive ability.

5. Prediction of drug toxicity with AI

Proportion of drug candidates being discarded during clinical trials due to unexpected adverse effects. Predicting drug toxicity during preclinical stages is a crucial step to reduce the failure rate and improve the efficiency of drug discovery. Traditional methods of predicting drug toxicity are limited by their reliance on small datasets and simplistic models. However, AI-based approaches have emerged as promising alternatives, leveraging large and diverse data sources, including chemical structures, biological pathways, and clinical data. By utilizing ML algorithms, AI-based approaches can improve the accuracy and efficiency of predicting the potentially toxic effects of new compounds, helping to mitigate risks associated with clinical trials, reduce drug development costs, and ultimately lead to better patient outcomes.

Recently, the use of AI-based computational models to forecast drug toxicity has grown in popularity [132]. Large drug and toxicity data sets have been analyzed in a number of studies using ML and DL algorithms, such as neural networks, to identify potential toxic effects during drug development. By identifying toxicities early on, these models can speed up the development of new drugs. Additionally, AI-based toxicity prediction models can prioritize compounds for testing and find new drug targets and toxicity mechanisms. AI toxicity prediction has been the subject of several reviews [132-137]. A single review is challenging due to the wide field of AI-based toxicity prediction's numerous toxicity properties. Recent AI-based toxicity prediction models and in-depth studies of toxicity properties are required to develop, optimize, and improve a model. For four important toxicity properties, recent ML and DL-based AI-based drug toxicity prediction methods are presented.

5.1. Cardiac side effect prediction

ML-based methods, including RF, SVM, NB, SVR, kNN, DT, GB (gradient boosting), PLS (partial least squares), and XGB, are commonly used in predicting hERG toxicity (a gene related to a potassium channel in the heart, and is used to evaluate cardiac side).

Venkatraman developed an ADMET prediction model using ECFP6-based RF models [138]. A total of 7889 compounds were assembled from 4 well-defined experimental assays with experimental hERG blocking bioactivities for training, where compounds with experimental values less than or equal to 10 μM were regarded as positive samples (4355 in total) and the others as negative samples (3534 in total). RF builds multiple DT on the data and merges them together to get hERG toxicity. This model achieved an 80% accuracy and 88% ROC-AUC.

Hsiao et al. [139] and Arab et al. [140] also used the RF model with high accuracy. Ogura et al. developed an SVM model using ECFP_4 structural fingerprints and 72 NSGA-II-selected descriptors, outperforming other predictors with 98.4% accuracy and 0.733 kappa statistic [141]. Konda et al. [142] generated hERG classification models with 2D descriptors using RF, SMO (sequential minimal optimization), and MLP (multilayer perceptron) algorithms, with their consensus model outperforming other predictors with 92% accuracy. DNN, ANN, RNN, CNN, GNN, GCNN, and GAT are used in developing hERG predictive models. Shan et al., [143] generated a directed message-passing neural network (DMPNN) model with moe206 descriptors that outperformed other models with an accuracy of 80% [143]. Zhang et al., [144] developed HergSPred, which outperformed other models, achieving an accuracy of 98.3%. Ryu et al. [145] developed DeepHit, which outperformed other tools in terms of accuracy, MCC, and SE (sensitivity). Wei et al. created Interpretable-ADMET, achieving the highest accuracy (91.9%) and ROC-AUC (78.2%) on 8672 compounds. ADMETLab 2.0 achieved the highest accuracy and ROC-AUC of 88.9% and 94.3%, respectively, on a large hERG data set containing 13,845 compounds, utilizing a multitask graph attention framework (MGAF) to predict ADMET properties.

5.2. LD_{50} prediction

In Toxicologists use the LD_{50} (median lethal dose) to determine a substance's toxicity. The dose needed to kill 50% of test animals in a particular period of time is referred to as the LD_{50} value of a chemical. The LD [146] is used as the first step in the drug screening process. Rats' acute oral toxicity was evaluated using LD_{50}. Interspecies variability and ethical issues render conventional LD_{50} testing obsolete [147]. Acute animal toxicity tests using tissue culture and in silico LD_{50} prediction are gradually being replaced [148]. The binary classification model divides substances into two categories: toxic (LD_{50} = 2000 mg/kg) and nontoxic (LD_{50} = 50 mg/kg).

Compounds are divided into multiple classes by the Globally Harmonized System of Classification and Labeling of Chemicals and the US Environmental Protection Agency [149]. Recent releases of the ADMET prediction programs FP-ADMET [138] and Interpretable-ADMET [150], in these both applications, LD_{50} prediction models are applied.

Ballabio et al. developed LD_{50} prediction models using the binary fingerprints of NB, N–Nearest Neighbors, Binned-Neighbors, and Extended Connectivity [151]. They found that their models were 84% sensitive and 81% specific for 8992 chemicals. An integrated QSAR model for 8448 compounds was created by Gadaleta et al., [152], by combining balanced RF, regression, aiQSAR, istkNN, SARpy, RF with hyperparameter tuning, and a general linear model [153]. The ideal model's RMSE (root-mean-squared error) was 0.477%, and its accuracy balance was above 70% for multiclass endpoints and 80% for binary endpoints.

Jain et al. [154] created a multitasking DL consensus model using RF, DNN, CNN, and GCNN on a dataset of 80,081 compounds that outperformed other models with RMSE of 0.65 and R^2 of 0.5. Using the same dataset of 7413 compounds, BTAMDL predicted LD_{50} with a higher degree of accuracy than MolGIN (RMSE = 0.557, R^2 = 0.662). Using the 7413-compound TopTox data set from the ECOTOX aquatic toxicity database, Karim et al., [155] proposed QuantitativeTox, a DL framework utilizing FFNN, CNN, GCNN, and baseline feature representations. This model outperformed others (R^2 = 0.687), and it was adopted.

5.3. Drug induced liver injury (DILI) prediction

Drug or chemical toxicity causes DILI [156]. It causes 32% of drug recalls, which worries researchers and doctors [157]. Predicting human DILI with *in vitro* or animal studies is difficult. RF, kNN, SVM, deep neural network (DNN), CNN, and GNN can predict compound properties from chemical structure. Recent reviews have examined AI methods for DILI in silico prediction [158-160].

Recent ML techniques, such as RF, LR, NB, SVM, kNN, AB, GDBT (gradient boosting decision trees), and ET (equivariant transformer) [138,161-165], have been applied to the development of accurate DILI prediction models. Mora et al. created an ensemble model based on QuBiLS-MAS Features and Shallow Learning using k-NN, MLP, RF, NB, SVM, LR (logistic regression), classification tree, Fisher's linear discriminant analysis (FLDA) [164], and Bayes network algorithms, which achieved an 84% accuracy rate over 10-fold cross-validation using a training set of 1075 compounds from a previous study [166]. Using a dataset of 2608 chemicals and SVM and hybrid quantum particle swarm optimization algorithms, Wang and Chen constructed five consensus models with an 80% accuracy rate [165].

ADMETLab 2.0 [167], FP-ADMET [138] and InterpretableADMET [150], are ADMET prediction software with over 79% accurate DILI prediction models. DL improves the accuracy of DILI prediction. Li et al.'s [168] DeepDILI model outperformed five ML algorithms and two advanced ensemble methods. Using Transcriptional Response Data and GNN, Hwang et al. [169] developed GLIT, a DILI prediction model that outperformed baseline models with 77.3% accuracy. With a multiview GNN approach, Ma et al. [170] increased accuracy from 78.8% to 81.4% and ROC-AUC from 86.6% to 88.8%. Using CNN algorithms and molecular fingerprint-embedded features, Nguyen-Vo et al. [171] developed a DILI prediction model with 89% accuracy and 96% ROC-AUC. CNN's ResNet18DNN achieved a 95.8% success rate on 1446 compounds [172]. These methods may reduce drug recalls and improve the accuracy of DILI prediction.

5.4. Carcinogenesis prediction

Potentially carcinogenic substances must be identified in order to prevent environmental cancers [173]. Many FDA-approved medications have been withdrawn due to their carcinogenic characteristics. Short-term biological studies and theoretical models have been tested to find such compounds. ML and DL techniques can be used to replace, scale back, and enhance animal studies.

Recently, a number of AI-based models and tools for compound carcinogenicity prediction were created [174-178]. Using hybrid neural networks (NN), Limbu and Dakshanamurthy developed carcinogenicity prediction models with an average accuracy of 74.3 and an average ROC-AUC of 80.1 [174]. Due to sparse data sets, DL models have low predictive accuracy. Wang et al. developed CapsCarcino, a new DL architecture, to address this problem [175]. On a set of external validation data, CapsCarcino had an average accuracy of 74.5 percent and an accuracy prediction rate of 83%. On sparse training data that was arbitrarily reduced to 20%, 40%, 60%, and 80% of the full training data, CapsCarcino performed better than other models. Li et al. developed the DeepCarc model, which has an average improvement rate of 37.0% and an accuracy of 75.4%, using 863 compounds and three descriptors [176]. Other reviews [173,179] cover additional models.

6. AI and gene editing technologies for developing gene therapies

With the growing accumulation of genomic and clinical data, data scientists face both challenges and opportunities when attempting to extract biologically or clinically relevant information from massive genotype and phenotype datasets. In genomics, AI-based technologies and data science techniques have been utilized effectively over the past two decades.

A significant amount of phenotypic information can be found in the clinical notes, discharge summaries, radiology, and pathology reports that make up about 80% of the unstructured data in EHRs [180]. Clinical NLP methods like cTAKES can parse semantic relationships and extract structured concepts from free text to extract this information. The accuracy of phenotyping has significantly improved with the use of NLP techniques in combination with structured and unstructured data. According to Liao et al.'s analysis of structured data, NLP can be added to it to increase sensitivity while preserving a high positive predictive value for a variety of illnesses, such as multiple sclerosis and inflammatory bowel disease [181].

A study on phenotyping rheumatoid arthritis using structured data (ICD codes and medication data) and clinical concepts derived from NLP shows how ML has been used to create phenotyping models. The SVM models outperformed rule-based techniques in accuracy [182], proving that a high-performing classifier can be built without the use of feature engineering. ML techniques are more scalable, work with less standardized datasets than rule-based approaches, and can capture more complex phenotypes.

However, manually labeled gold standard training and test datasets are crucial for developing and validating supervised ML models. However, creating them requires considerable time and expertise. To address this, unsupervised learning techniques have been proposed to generate patient clusters for specific medical conditions without human supervision. Ho et al. used the "Limestone" non-negative tensor factorization technique to automatically generate multiple phenotype candidates without predefined definitions [183]. A medical professional determined that only 40 of the top 50 candidates are necessary for higher predictive accuracy of patients at risk of heart failure. Two upgraded variations of Limestone, Marble, and Granite, showed improved performance [184,185]. Numerous phenotyping methods have been reported in line with the use of DL approaches, including Gehrmann et al., [186] and Yang et al., [187] use of discharge summaries or clinicians' notes, and Miotto et al., [188] de-noising auto-encoders for auto-encoding.

In genomics research, DL has emerged as a prominent class of algorithms owing to its capability to effectively handle large datasets with high dimensionality. A plethora of DL-based models have been developed and utilized for designing gRNA (guide RNA), incorporating features from both sequence and secondary structure data. Additionally, transfer learning, a ground-breaking innovation in computer vision, has been used in genomics research as well to take advantage of trained models and only need small sample sizes [189]. Moreover, BERT models have been specially created for NLP tasks in the clinical domain [190]. ClinicalBERT and Discharge Summary BERT were developed by Alsentzer et al., [191] and pre-trained on millions of clinical notes from the MIMIC-III database [192,193]. These MIMIC notes were split into sections followed by sentences extraction as input into the model, which captured contextual relationships between words bidirectionally. These models were pre-trained and fine-tuned for downstream tasks, outperforming BERT and BioBERT on three clinical NLP tasks. In order to predict hospital readmission, Huang et al. also created ClinicalBERT, but they also emphasized the drawbacks of using data from a single healthcare institution. Retraining on bigger databases of clinical notes is advised as a result for better performance. These advancements demonstrate how ML and its related fields can enhance genomics and clinical research.

7. AI-based modeling for personalized drug dosing

Traditionally, clinical practice has been based on the concept of "one therapy fits all'. However, drug molecules may undergo different metabolic activities in different patients. For example, a drug that works well for a group of people may not be as effective or may have adverse side effects for others. These differences in drug metabolism are mostly attributed to the differences in the genetic profile of individuals. Thus, a more futuristic approach is the personalized treatment also known as precision medicine, where patients are treated based on their genetic profile. The target is to maximize treatment outcomes while minimizing adverse effects per individual. Thus, different therapies and doses are customized per individual (or per group of patients that share similar genome profiles). AI has fostered considerable improvements in the development of personalized medicine [194]. For example, the AI-derived platform, CURATE.AI, predicts the optimal dosing along with the treatment outcomes based on the patients' individual data. It gen-

erates a profile for each patient using their own medical records, and it dynamically recalibrates the predicted profile over time based on the progression or recession of the disease. CURATE.AI can optimize doses of not only single drugs, but also combinations of drugs [195,196]. This is helpful given that, nowadays, therapies are becoming more sophisticated with emphasis on combination (or multimodal) treatments. These involve more than one drug or treatment offered either simultaneously or sequentially. Combination therapy is proven to have more efficacy compared to single drug regimen especially in the treatment of complex diseases like cancer [197,198]. To predict the efficacy of a chosen treatment, Kureshi et al. developed an AI decision tree to establish a link between the characteristics of the patient and the tumor response in NSCLC [199]. They used four classifiers (histology, mutation in epidermal growth factor receptor, targeted drugs, and smoking habits) for predicting the response of NSCLC patients to the EGFR tyrosine kinase inhibitors. The method showed 76.6% accuracy and it can support the clinician in choosing the correct treatment for NSCLC patients. One of the drawbacks of the study is the small training set used (n = 355). This resulted in the omission of rare patterns such as duplication, deletions, insertions, and point mutations. Using a larger training set could further improve the predictive accuracy of this decision support model. The 'IBM Watson for oncology' software made a large impact on personalized treatment plans for cancer patients [200]. The software is trained on thousands of clinical and health records of cancer patients from the medical journals, textbooks, and literature curated by Memorial Sloan Kettering. This software makes accurate diagnoses and treatment recommendations by identifying related cases from databases of worldwide clinical trials (http://www.clinicaltrials.gov) [201].

8. The role of AI in rare disease research

Rare diseases (RDs) are a significant health issue that affects almost 1 in 10 individuals in US [202]. Despite their prevalence, the diagnosis of RDs is often challenging due to the complexity of symptoms and the rarity of the conditions. The delay in diagnosis can be as long as 7 years, leading to significant delays in treatment and management [203]. Hence, there is a need for new approaches to enhance the diagnosis and treatment of RDs. AI has the potential to transform the diagnosis and management of rare diseases [204-207] based on NB, RF, XGBoost, CNN, AE (autoencoder), RNN, GAN (generative adversarial network), etc. Fernández et al. developed a deep DL-based approach to detect tubers in selected MRI (magnetic resonance imaging) images for the diagnosis of tuberous sclerosis complex (TSC) [204]. This model adopts a unique InceptionV3 CNN architecture to recognize whether an MRI image has tubers in it or not, showing promising performance (accuracy: 95%) in the detection of a rare neurological disorder. Founta introduced a semi-automated preprocessing gene selection methodology to identify causal amyotrophic lateral sclerosis (ALS) genes [205], with which they developed a classifier based on XGBoost and RF to diagnose ALS and its specific subtypes. This methodology achieved 88.89% accuracy for the classification of sporadic ALS motor neuron samples. Additionally, AI-based PET is a promising tool for early detection and diagnosis of RDs [208].

However, the implementation of AI in healthcare requires careful consideration of ethical, legal, and social implications [209]. AI medical devices must be developed with the active involvement of patient advocacy groups to ensure that the technology is designed to meet the specific needs of rare disease patients. The datasets used to train these algorithms must be diverse and augmented to ensure that they represent the end-user population accurately. Furthermore, the safety and effectiveness of AI-based medical devices (AIMDs) must be thoroughly evaluated to avoid potential harm to patients [210]. AIMDs must be RD-aware at every stage of their conceptualization and life cycle to avoid potential harm and unsustainable deployment of AIMDs into clinical practice. This requires a multidisciplinary approach involving clinicians, computer scientists, and patient advocacy groups.

In general, AI-based technologies offer promising solutions to improve the diagnosis and management of rare diseases. However, the ethical, legal, and social implications of AI in healthcare must be carefully considered to ensure the safety and effectiveness of AIMDs. With careful consideration and collaboration, AI has the potential to revolutionize the diagnosis and treatment of rare diseases, leading to improved patient outcomes and a better quality of life for those affected by RDs.

9. Conclusion

The use of AI technology in drug design has grown rapidly due to its predictive ability and accuracy. This review highlights the numerous applications of AI in all phases of drug development, from disease diagnosis to post-marketing analysis. AI helps in the early prediction of diseases, the development of personalized medicine, optimization of drug doses, and the prediction of treatment outcomes. Additionally, AI assists in target and lead identification through the prediction of protein structures and biological activities of small molecules. AI technology can also predict drug-like properties and off-target effects of new compounds, reducing the need for experimental validation. Furthermore, AI-driven approaches improve patient stratification, recruitment, monitoring, and follow-ups in clinical trials, and can even assist in FDA approvals and pharmacovigilance. The integration of AI in drug design has resulted in faster drug discovery, cost savings, reduced resource and manpower usage, and decreased attrition rates in clinical trials. Additionally, AI helps to minimize the use of *in vivo* bioassays, reducing animal sacrifice. AI has far-reaching applications beyond medicine, including healthcare management, surgeries, mRNA vaccination, preventive treatments, and nutrigenomics. However, it is important to note that AI models are meant to complement human intelligence, not replace it. AI models may have comparable or better predictive ability than human researchers, but they still lack human intuition. Predictions made by AI machines must be verified by humans, as AI models can provide false positive and false negative results, compromising the sensitivity and specificity of the model. Additionally, resource sustainability needs holistic solutions like cost-aware cross-layer co-design, integrating hardware, algorithms, and models for efficient exploration of resource-sustainable configurations. Consensus-based distributed learning is suggested to fully utilize existing and future computing infrastructures, incorporating Internet-of-Things devices and edge servers for data sharing while ensuring privacy. Stable infrastructures with AI-enhanced resource allocation are recommended, involving dedicated healthcare AI infrastructures compliant with evolving government regulations. Lastly, interpretable self-supervised learning is proposed to address the sustainability issue in domain expertise, enhancing trust by extracting clinically useful features and providing human-interpretable evidence in healthcare applications. There are numerous challenges associated with AI, including the explainability of models, the quality and suitability of data used to train models, avoiding bias and overfitting, resource sustainability and more. It is crucial to remain aware of the limitations and risks associated with AI technology. Opportunities for improvement in AI technology include minimizing dependence on supercomputing power, addressing ethical concerns surrounding data collection, and implementing AI in a controlled manner in the healthcare sector to limit negative consequences. It is possible that the future of AI-assisted drug discovery lies in developing a virtual human with complete complexity, allowing for accurate predictions of all possible interactions between molecules and exploring all therapeutic potentials and adverse side effects.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work was supported by the National Key R&D Program of China (2023YFF1205103), National Natural Science Foundation of China (81925034, 22237005), the Key Research and Development Program of Ningxia Hui Autonomous Region (2022CMG01002), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-007), the innovative research team of high-level local universities in Shanghai (SHSMU-ZDCX20212700).

References

[1] I.V. Hinkson, B. Madej, E.A. Stahlberg, Accelerating therapeutics for opportunities in medicine: A paradigm shift in drug discovery, Front. Pharmacol. 11 (2020) 770.
[2] R. Barker, A flexible blueprint for the future of drug development, Lancet 375 (2010) 357–359.
[3] D. Sun, W. Gao, H. Hu, et al., Why 90% of clinical drug development fails and how to improve it? Acta. Pharm. Sin. B 12 (2022) 3049–3062.
[4] H. Dowden, J. Munro, Trends in clinical success rates and therapeutic focus, Nat. Rev. Drug Discov. 18 (2019) 495–496.
[5] P. Hassanzadeh, F. Atyabi, R. Dinarvand, The significance of artificial intelligence in drug delivery system design, Adv. Drug Deliv. Rev. 151-152 (2019) 169–190.
[6] J.D. Bolcer, R.B. Hermann, The development of computational chemistry in the United States, Rev. Comput. Chem. (1994) 1–63.
[7] F. Bianconi, M. Filippucci, Digital Wood Design: Innovative Techniques of Representation in Architectural Design, Springer, 2019 Vol. 24.
[8] Y. Ding, J.H. Sohn, M.G. Kawczynski, et al., A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain, Radiology 290 (2019) 456–464.
[9] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks, Pattern Anal. Appl. 24 (2021) 1207–1220.
[10] F. Gentile, J.C. Yaacoub, J. Gleave, et al., Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking, Nat. Protoc. 17 (2022) 672–697.
[11] K.A. Carpenter, D.S. Cohen, J.T. Jarrell, et al., Deep learning and virtual drug screening, Future Med. Chem. 10 (2018) 2557–2567.
[12] M. Bule, N. Jalalimanesh, Z. Bayrami, et al., The rise of deep learning and transformations in bioactivity prediction power of molecular modeling tools, Chem. Biol. Drug Des. 98 (2021) 954–967.
[13] C. Berzuini, R. Bellazzi, S. Quaglini, et al., Bayesian networks for patient monitoring, Artif. Intell. Med. 4 (1992) 243–260.
[14] S. Kapsiani, B.J. Howlin, Random forest classification for predicting lifespan-extending chemical compounds, Sci. Rep. 11 (2021) 1–13.
[15] K. Heikamp, J. Bajorath, Support vector machines for drug discovery, Expert Opin. Drug Discov. 9 (2014) 93–104.
[16] F. Zhong, J. Xing, X. Li, et al., Artificial intelligence in drug design, Sci. China Life. Sci. 61 (2018) 1191–1204.
[17] N. Brown, P. Ertl, R. Lewis, et al., Artificial intelligence in chemistry and drug design, J. Comput. Aided Mol. Des. 34 (2020) 709–715.
[18] C. Provenzano, M. Cappella, R. Valaperta, et al., CRISPR/Cas9-mediated deletion of CTG expansions recovers normal phenotype in myogenic cells derived from myotonic dystrophy 1 patients, Mol. Ther. Nucl. Acids 9 (2017) 337–348.
[19] H.-C. Yi, Z.-H. You, X. Zhou, et al., ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, Mol. Ther. Nucl. Acids 17 (2019) 1–9.
[20] H. Yi, Z.H. You, X. Zhou, L. Cheng, X. Li, T.H. Jiang, ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, Mol. Ther. Nucl. Acids 17 (2019) 9.
[21] C. Shen, J. Ding, Z. Wang, et al., From machine learning to deep learning: Advances in scoring functions for protein–ligand docking, WIREs Comput. Mol. Sci. 10 (2020) e1429.
[22] Y. Shen, T. Liu, J. Chen, et al., Harnessing artificial intelligence to optimize long-term maintenance dosing for antiretroviral-naive adults with HIV-1 infection, Adv. Ther. (Weinh) 3 (2020) 1900114.
[23] M. Zhang, Q. Su, Y. Lu, M. Zhao, Application of machine learning approaches for protein-protein Interactions prediction, Med. Chem. 13 (2017) 506–514.
[24] F. Yu, H.H. Ip, Semantic content analysis and annotation of histological images, Comput. Biol. Med. 38 (2008) 635–649.
[25] G.B. Goh, N.O. Hodas, A. Vishnu, Deep learning for computational chemistry, J. Comput. Chem. 38 (2017) 1291–1307.
[26] M. Singer, C.S. Deutschman, C.W. Seymour, et al., The third international consensus definitions for sepsis and septic shock (Sepsis-3), JAMA 315 (2016) 801–810.
[27] M.P. Sendak, W. Ratliff, D. Sarro, et al., Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study, JMIR Med. Inform. 8 (2020) e15182.
[28] G. Husabø, I.L. Teig, J.C. Frich, et al., Promoting leadership and quality improvement through external inspections of management of sepsis in Norwegian hospitals: A focus group study, BMJ Open 10 (2020) e041997.
[29] N. Alballa, I. Al-Turaiki, Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review, Inform. Med. Unlocked 24 (2021) 100564.
[30] L. Yu, X. Shi, X. Liu, et al., Artificial intelligence systems for diagnosis and clinical classification of COVID-19, Front. Microbiol. 12 (2021) 729455.
[31] A.A. Castro, T.D. Antonio, E.C. Martínez, et al., Usefulness of chest X-rays for evaluating prognosis in patients with COVID-19, Radiologia 63 (2021) 476–483.
[32] W. Xie, F. Wang, Y. Li, et al., Advances and challenges in de novo drug design using three-dimensional deep generative models, J. Chem. Inf. Model 62 (2022) 2269–2279.
[33] W.T. Li, J. Ma, N. Shende, et al., Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis, BMC Med. Inform. Decis. Mak. 20 (2020) 1–13.
[34] L. Capoferri, A. Lodola, S. Rivara, et al., Quantum mechanics/molecular mechanics modeling of covalent addition between EGFR-cysteine 797 and N-(4-anilinoquinazolin-6-yl) acrylamide, J. Chem. Inf. Model 55 (2015) 589–599.
[35] A. Chatzigoulas, Z. Cournia, Rational design of allosteric modulators: Challenges and successes, WIREs Comput. Mol. Sci. 11 (2021) e1529.
[36] Q. Zhang, Y. Chen, D. Ni, et al., Targeting a cryptic allosteric site of SIRT6 with small-molecule inhibitors that inhibit the migration of pancreatic cancer cells, Acta. Pharm. Sin. B 12 (2022) 876–889.
[37] S. Spänig, A. Emberger-Klein, J.-P. Sowa, et al., The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes, Artif. Intell. Med. 100 (2019) 101706.
[38] V. Gulshan, L. Peng, M. Coram, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (2016) 2402–2410.
[39] A.A. Van Der Heijden, M.D. Abramoff, F. Verbraak, et al., Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System, Acta. Ophthalmol. 96 (2018) 63–68.
[40] X. Liu, K. Chen, T. Wu, et al., Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease, Transl. Res. 194 (2018) 56–67.
[41] S. Basheer, S. Bhatia, S.B. Sakri, Computational modeling of dementia prediction using deep neural network: Analysis on OASIS dataset, IEEE Access 9 (2021) 42449–42462.
[42] T. Jo, K. Nho, A.J. Saykin, Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data, Front. Aging Neurosci. 11 (2019) 220.
[43] X.H. Ma, Z. Shi, C. Tan, et al., In-silico approaches to multi-target drug discovery: Computer aided multi-target drug design, multi-target virtual screening, Pharm. Res. 27 (2010) 739–749.
[44] I. Beheshti, H. Demirel, H. Matsuda, et al., Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm, Comput. Biol. Med. 83 (2017) 109–119.
[45] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.
[46] A. Albayrak, G. Bilgin, in: Proceedings of the 2016 IEEE 17th International symposium on computational intelligence and informatics (CINTI), 2016, pp. 000335–000340.
[47] J.L. Causey, J. Zhang, S. Ma, et al., Highly accurate model for prediction of lung nodule malignancy with CT scans, Sci. Rep. 8 (2018) 9286.
[48] I. Shiri, H. Maleki, G. Hajianfar, et al., Next-generation radiogenomics sequencing for prediction of EGFR and KRAS mutation status in NSCLC patients using multimodal imaging and machine learning algorithms, Mol. Imaging Biol. 22 (2020) 1132–1148.
[49] C.L. Srinidhi, O. Ciga, A.L. Martel, Deep neural network models for computational histopathology: A survey, Med. Image Anal. 67 (2021) 101813.
[50] Ş. Öztürk, B. Akdemir, HIC-net: A deep convolutional neural network model for classification of histopathological breast images, Comput. Electr. Eng. 76 (2019) 299–310.
[51] Z. Hameed, S. Zahia, B. Garcia-Zapirain, et al., Breast cancer histopathology image classification using an ensemble of deep learning models, Sensors 20 (2020) 4373.
[52] B.S. Yadav, V. Tripathi, Recent advances in the system biology-based target identification and drug discovery, Curr. Top. Med. Chem. 18 (2018) 1737–1744.
[53] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Expert Opin. Drug Discov. 5 (2010) 633–654.
[54] A. Assaiya, A.P. Burada, S. Dhingra, et al., An overview of the recent advances in cryo-electron microscopy for life sciences, Emerg. Top. Life Sci. 5 (2021) 151–168.
[55] S.G. Wolf, E. Shimoni, M. Elbaum, et al., STEM tomography in biology, in: Cellular Imaging: Electron Tomography and Related Techniques, 2018, pp. 33–60.
[56] E.D. Zhong, T. Bepler, B. Berger, et al., CryoDRGN: Reconstruction of heterogeneous cryo-EM structures using neural networks, Nat. Methods 18 (2021) 176–185.
[57] L.F. Kinman, B.M. Powell, E.D. Zhong, et al., Uncovering structural ensembles from single-particle cryo-EM data using cryoDRGN, Nat. Protoc. 18 (2023) 319–339.
[58] J. Yang, Z. Gao, X. Ren, et al., DeepDigest: Prediction of protein proteolytic digestion with deep learning, Anal. Chem. 93 (2021) 6094–6103.
[59] B. Sun, P. Smialowski, T. Straub, et al., Investigation and highly accurate prediction of missed tryptic cleavages by deep learning, J. Proteome Res. 20 (2021) 3749–3757.
[60] K. Murata, M. Wolf, Cryo-electron microscopy for structural analysis of dynamic biological macromolecules, Biochimica et Biophysica Acta 1862 (2018) 324–334.
[61] B. Webb, A. Sali, Comparative protein structure modeling using MODELLER, Curr. Protoc. Bioinformat. 54 (2016) 5.6. 1-5.6. 37.
[62] J. Jumper, R. Evans, A. Pritzel, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589.

[63] K. Tunyasuvunakool, J. Adler, Z. Wu, et al., Highly accurate protein structure prediction for the human proteome, Nature 596 (2021) 590–596.
[64] G.R. Buel, K.J. Walters, Can AlphaFold2 predict the impact of missense mutations on structure? Nat. Struct. Mol. Biol. 29 (2022) 1–2.
[65] A. Perrakis, T.K. Sixma, AI revolutions in biology: The joys and perils of AlphaFold, EMBO Rep. 22 (2021) e54046.
[66] H. Tian, X. Jiang, P. Tao, PASSer: Prediction of allosteric sites server, Mach. Learn. Sci. Technol. 2 (2021) 035015.
[67] W. Huang, S. Lu, Z. Huang, et al., Allosite: A method for predicting allosteric sites, Bioinformatics 29 (2013) 2357–2359.
[68] J.G. Greener, M.J. Sternberg, AlloPred: Prediction of allosteric pockets on proteins using normal mode perturbation analysis, BMC Bioinformat. 16 (2015) 1–7.
[69] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural. Process Lett. 9 (1999) 293–300.
[70] A. Liaw, M. Wiener, Classification and regression by randomForest, R news 2 (2002) 18–22.
[71] T. Chen, C. Guestrin, in: Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[72] A.H. Basit, W.A. Abbasi, A. Asif, et al., Training host-pathogen protein–protein interaction predictors, J. Bioinform. Comput. Biol. 16 (2018) 1850014.
[73] K. Li, S. Zhang, D. Yan, et al., Prediction of hot spots in protein–DNA binding interfaces based on supervised isometric feature mapping and extreme gradient boosting, BMC Bioinformat. 21 (2020) 1–10.
[74] R.E. Amaro, Will the real cryptic pocket please stand out? Biophys. J. 116 (2019) 753–754.
[75] C.R. Knoverek, G.K. Amarasinghe, G.R. Bowman, Advanced methods for accessing protein shape-shifting present new therapeutic opportunities, Trends Biochem. Sci. 44 (2019) 351–364.
[76] C.C. Witt, Y. Ono, E. Puschmann, et al., Induction and myofibrillar targeting of CARP, and suppression of the Nkx2. 5 pathway in the MDM mouse with impaired titin-based signaling, J. Mol. Biol. 336 (2004) 145–154.
[77] M.A. Cruz, T.E. Frederick, U.L. Mallimadugula, et al., A cryptic pocket in Ebola VP35 allosterically controls RNA binding, Nat. Commun. 13 (2022) 2269.
[78] A. Meller, M.D. Ward, J.H. Borowsky, et al., Predicting the locations of cryptic pockets from single protein structures using the PocketMiner graph neural network, Biophys. J. 122 (2023) 445a.
[79] S.A. Hollingsworth, B. Kelly, C. Valant, et al., Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic GPCRs, Nat. Commun. 10 (2019) 3289.
[80] P. Cimermancic, P. Weinkam, T.J. Rettenmaier, et al., CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites, J. Mol. Biol. 428 (2016) 709–719.
[81] Q. Li, T. Cheng, Y. Wang, et al., PubChem as a public resource for drug discovery, Drug Discov. Today 15 (2010) 1052–1057.
[82] B. Chen, D.J. Wild, PubChem BioAssays as a data source for predictive models, J. Mol. Graph. Modell. 28 (2010) 420–426.
[83] J. Lamb, The Connectivity Map: A new tool for biomedical research, Nat. Rev. Cancer 7 (2007) 54–60.
[84] J. Lamb, E.D. Crawford, D. Peck, et al., The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease, Science 313 (2006) 1929–1935.
[85] I. Kozlovskii, P. Popov, Structure-based deep learning for binding site detection in nucleic acid macromolecules, NAR Genom. Bioinform. 3 (2021) lqab111.
[86] I. Kozlovskii, P. Popov, Spatiotemporal identification of druggable binding sites using deep learning, Commun. Biol. 3 (2020) 618.
[87] I. Kozlovskii, P. Popov, Protein–peptide binding site detection using 3D convolutional neural networks, J. Chem. Inf. Model 61 (2021) 3814–3823.
[88] X. Zeng, S. Zhu, W. Lu, et al., Target identification among known drugs by deep learning from heterogeneous networks, Chem. Sci. 11 (2020) 1775–1797.
[89] P. Mamoshina, M. Volosnikova, I.V. Ozerov, et al., Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification, Front. Genet. 9 (2018) 242.
[90] S. Dara, S. Dhamercherla, S.S. Jadav, et al., Machine learning in drug discovery: A review, Artif. Intell. Rev. 55 (2022) 1947–1999.
[91] W. Feng, L. Wang, Z. Lin, et al., Generation of 3D Molecules in Pockets via Language Model, Nat. Mach. Intell. 6 (2024) 62–73.
[92] A. Smith, Screening for drug discovery: The leading question, Nature 418 (2002) 453–455.
[93] M. Butkiewicz, Y. Wang, S.H. Bryant, et al., High-throughput screening assay datasets from the pubchem database, Chem. Inform. 3 (2017) 1.
[94] N. Berdigaliyev, M. Aljofan, An overview of drug discovery and development, Future Med. Chem. 12 (2020) 939–947.
[95] N.S. Pagadala, K. Syed, J. Tuszynski, Software for molecular docking: A review, Biophys. Rev. 9 (2017) 91–102.
[96] N. Brooijmans, I.D. Kuntz, Molecular recognition and docking algorithms, Ann. Rev. Biophys. Biomol. Struct. 32 (2003) 335–373.
[97] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inform. 29 (2010) 476–488.
[98] K.A. Carpenter, X. Huang, Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: A review, Curr. Pharm. Des. 24 (2018) 3347–3358.
[99] L. Wang, X. Le, L. Li, et al., Discovering new agents active against methicillin-resistant *Staphylococcus aureus* with ligand-based approaches, J. Chem. Inf. Model 54 (2014) 3186–3197.
[100] W. Lian, J. Fang, C. Li, et al., Discovery of Influenza A virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models, Mol. Divers. 20 (2016) 439–451.
[101] M. Mochizuki, S.D. Suzuki, K. Yanagisawa, et al., QEX: Target-specific druglikeness filter enhances ligand-based virtual screening, Mol. Divers. 23 (2019) 11–18.
[102] Y. Li, L. Wang, Z. Liu, et al., Predicting selective liver X receptor β agonists using multiple machine learning methods, Mol. Biosyst. 11 (2015) 1241–1250.
[103] L. Zhang, A. Sedykh, A. Tripathi, et al., Identification of putative estrogen receptor–mediated endocrine disrupting chemicals using QSAR-and structure-based virtual screening approaches, Toxicol. Appl. Pharmacol. 272 (2013) 67–76.
[104] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[105] S. Chandra, J. Pandey, A.K. Tamrakar, et al., Multiple machine learning based descriptive and predictive workflow for the identification of potential PTP1B inhibitors, J. Mol. Graph. Modell. 71 (2017) 242–256.
[106] A.L. Deshmukh, S. Chandra, D.K. Singh, et al., Identification of human flap endonuclease 1 (FEN1) inhibitors using a machine learning based consensus virtual screening, Mol. Biosyst. 13 (2017) 1630–1639.
[107] H. Baba, J.-i. Takahara, H. Mamitsuka, In silico predictions of human skin permeability using nonlinear quantitative structure–property relationship models, Pharm. Res. 32 (2015) 2360–2371.
[108] M. AlQuraishi, End-to-end differentiable learning of protein structure, Cell. Syst. 8 (2019) 292–301 e293.
[109] K. Lee, M. Lee, D. Kim, Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server, BMC Bioinformat. 18 (2017) 75–86.
[110] C.F. Matta, A.A. Arabi, Electron-density descriptors as predictors in quantitative structure–activity/property relationships and drug design, Future Med. Chem. 3 (2011) 969–994.
[111] A.M. Osman, A.A. Arabi, Quantum and classical evaluations of carboxylic acid bioisosteres: From capped moieties to a drug molecule, ACS Omega 8 (2022) 588–598.
[112] A.R. Melge, S. Parate, K. Pavithran, et al., Discovery of anticancer hybrid molecules by supervised machine learning models and in vitro validation in drug resistant chronic myeloid leukemia cells, J. Chem. Inf. Model 62 (2022) 1126–1146.
[113] G. Dhamodharan, C.G. Mohan, Machine learning models for predicting the activity of AChE and BACE1 dual inhibitors for the treatment of Alzheimer's disease, Mol. Divers. (2022) 1–17.
[114] J. Hardy, D.J. Selkoe, The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics, Science (1979) 297 (2002) 353–356.
[115] M. Rapoport, H.N. Dawson, L.I. Binder, et al., Tau is essential to β-amyloid-induced neurotoxicity, Proc. Natl. Acad. Sci. 99 (2002) 6364–6369.
[116] N.J. Woolf, The critical role of cholinergic basal forebrain neurons in morphological change and memory encoding: A hypothesis, Neurobiol. Learn. Mem. 66 (1996) 258–266.
[117] M.P. Mazanetz, P.M. Fischer, Untangling tau hyperphosphorylation in drug design for neurodegenerative diseases, Nat. Rev. Drug Discov. 6 (2007) 464–479.
[118] W. Sun, H.Y. Qureshi, P.W. Cafferty, et al., Glycogen synthase kinase-3β is complexed with tau protein in brain microtubules, J. Biol. Chem. 277 (2002) 11933–11940.
[119] A. Arif, Extraneuronal activities and regulatory mechanisms of the atypical cyclin-dependent kinase Cdk5, Biochem. Pharmacol. 84 (2012) 985–993.
[120] J. Fang, Y. Li, R. Liu, et al., Discovery of multitarget-directed ligands against Alzheimer's disease through systematic prediction of chemical–protein interactions, J. Chem. Inf. Model 55 (2015) 149–164.
[121] C. Gorgulla, A. Boeszoermenyi, Z.-F. Wang, et al., An open-source drug discovery platform enables ultra-large virtual screens, Nature 580 (2020) 663–668.
[122] B.J. Bender, S. Gahbauer, A. Luttens, et al., A practical guide to large-scale docking, Nat. Protoc. 16 (2021) 4799–4832.
[123] J. Lyu, S. Wang, T.E. Balius, et al., Ultra-large library docking for discovering new chemotypes, Nature 566 (2019) 224–229.
[124] W.L. Jorgensen, The many roles of computation in drug discovery, Science 303 (2004) 1813–1818.
[125] Y. Li, M.A. Rezaei, C. Li, et al., in: Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 303–310.
[126] J. Lim, S. Ryu, K. Park, et al., Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation, J. Chem. Inf. Model 59 (2019) 3981–3988.
[127] M. Jiang, Z. Li, S. Zhang, et al., Drug–target affinity prediction using graph neural network and contact maps, RSC Adv. 10 (2020) 20701–20712.
[128] M. Rudrapal, S.J. Khairnar, A.G. Jadhav, Drug repurposing (DR): An emerging approach in drug discovery, Drug Repurposing-Hypothesis, Molecular Aspects and Therapeutic Applications (2020) 10.
[129] S.H. Sleigh, C.L. Barton, Repurposing strategies for therapeutics, Pharmaceut. Med. 24 (2010) 151–159.
[130] R.K. Harrison, Phase II and phase III failures: 2013–2015, Nat. Rev. Drug Discov. 15 (2016) 817–818.
[131] D. Reker, T. Rodrigues, P. Schneider, et al., Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus, Proc. Natl. Acad. Sci. 111 (2014) 4067–4072.
[132] X. Yang, Y. Wang, R. Byrne, et al., Concepts of artificial intelligence for computer-assisted drug discovery, Chem. Rev. 119 (2019) 10520–10594.
[133] A.H. Vo, T.R. Van Vleet, R.R. Gupta, et al., An overview of machine learning and big data for drug toxicity evaluation, Chem. Res. Toxicol. 33 (2020) 20–37.
[134] J. Liu, W. Guo, F. Dong, et al., machine learning for predicting organ toxicity, in: H. Hong (Ed.), Machine Learning and Deep Learning in Computational Toxicology, Springer International Publishing, Cham, 2023, pp. 519–537.

[135] A.O. Basile, A. Yahi, N.P. Tatonetti, Artificial intelligence for drug toxicity and safety, Trends Pharmacol. Sci. 40 (2019) 624–635.
[136] E. Pérez Santín, R. Rodríguez Solana, M. González García, et al., Toxicity prediction based on artificial intelligence: A multidisciplinary overview, WIREs Comput. Mol. Sci. 11 (2021) e1516.
[137] A. Bassan, V.M. Alves, A. Amberg, et al., In silico approaches in organ toxicity hazard assessment: Current status and future needs in predicting liver toxicity, Comput. Toxicol. 20 (2021) 100187.
[138] V. Venkatraman, FP-ADMET: A compendium of fingerprint-based ADMET prediction models, J. Cheminform. 13 (2021) 1–12.
[139] Y. Hsiao, B.-H. Su, Y.J. Tseng, Current development of integrated web servers for preclinical safety and pharmacokinetics assessments in drug development, Brief Bioinformat. 22 (2021) bbaa160.
[140] I. Arab, K. Barakat, ToxTree: Descriptor-based machine learning models for both hERG and Nav1. 5 cardiotoxicity liability predictions, arXiv preprint (2021) arXiv:2112.13467 (Accessed Dec 27, 2021).
[141] K. Ogura, T. Sato, H. Yuki, et al., Support vector machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II, Sci. Rep. 9 (2019) 1–12.
[142] L.S.K. Konda, S.K. Praba, R. Kristam, hERG liability classification models using machine learning techniques, Comput. Toxicol. 12 (2019) 100089.
[143] M. Shan, C. Jiang, J. Chen, et al., Predicting hERG channel blockers with directed message passing neural networks, RSC Adv. 12 (2022) 3423–3430.
[144] X. Zhang, J. Mao, M. Wei, et al., HergSPred: Accurate classification of hERG blockers/nonblockers with machine-learning models, J. Chem. Inf. Model 62 (2022) 1830–1839.
[145] J.Y. Ryu, M.Y. Lee, J.H. Lee, et al., DeepHIT: A deep learning framework for prediction of hERG-induced cardiotoxicity, Bioinformatics 36 (2020) 3049–3055.
[146] R. De Jesus, N. Vicuña-Fernández, A. Osorio, et al., Determination of medium lethal dose (LD50) and acute toxicity of formulation Cytoreg®, an ionic mixture of strong and weak acids, Latin Am. J. Dev. 3 (2021) 1121–1126.
[147] V. Bhat, J. Chatterjee, The use of in silico tools for the toxicity prediction of potential inhibitors of SARS-CoV-2, Alternat. Lab. Anim. 49 (2021) 22–32.
[148] K. Morris-Schaffer, M.J. McCoy, A review of the LD50 and its current role in hazard communication, ACS Chem. Health Saf. 28 (2020) 25–33.
[149] J. Kramer, Label Review Manual Chapter 7: Precautionary Statements Label Review Manual, (2014).
[150] Y. Wei, S. Li, Z. Li, et al., Interpretable-ADMET: A web service for ADMET prediction and optimization based on deep neural representation, Bioinformatics 38 (2022) 2863–2871.
[151] D. Ballabio, F. Grisoni, V. Consonni, et al., Integrated QSAR models to predict acute oral systemic toxicity, Mol. Inform. 38 (2019) 1800124.
[152] K. Vukovic, D. Gadaleta, E. Benfenati, Methodology of aiQSAR: A group-specific approach to QSAR modelling, J. Cheminform. 11 (2019) 1–9.
[153] D. Gadaleta, K. Vuković, C. Toma, et al., SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data, J. Cheminform. 11 (2019) 1–16.
[154] S. Jain, V.B. Siramshetty, V.M. Alves, et al., Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods, J. Chem. Inf. Model 61 (2021) 653–663.
[155] A. Karim, V. Riahi, A. Mishra, et al., Quantitative toxicity prediction via meta ensembling of multitask deep learning models, ACS Omega 6 (2021) 12306–12317.
[156] R.J. Andrade, N. Chalasani, E.S. Björnsson, et al., Drug-induced liver injury, Nat. Rev. Dis. Prim. 5 (2019) 58.
[157] S. Babai, L. Auclert, H. Le-Louët, Safety data and withdrawal of hepatotoxic drugs, Therapies 76 (2021) 715–723.
[158] O.J. Béquignon, G. Pawar, B. van de Water, et al., Computational approaches for drug-induced liver injury (DILI) prediction: State of the art and challenges, Syst. Med. 2 (2021) 308–329.
[159] V.M. Lauschke, Toxicogenomics of drug induced liver injury–from mechanistic understanding to early prediction, Drug Metab. Rev. 53 (2021) 245–252.
[160] J. Liu, W. Guo, S. Sakkiah, et al., Machine learning models for predicting liver toxicity, in: In Silico Methods for Predicting Drug Toxicity, 2022, pp. 393–415.
[161] M.G. de Lomana, F. Svensson, A. Volkamer, et al., Consideration of predicted small-molecule metabolites in computational toxicology, Digit. Discov. 1 (2022) 158–172.
[162] X. Liu, D. Zheng, Y. Zhong, et al., Machine-learning prediction of oral drug-induced liver injury (DILI) via multiple features and endpoints, Biomed. Res. Int. 2020 (2020) 4795140.
[163] Y. Wang, Q. Xiao, P. Chen, et al., In silico prediction of drug-induced liver injury based on ensemble classifier method, Int. J. Mol. Sci. 20 (2019) 4106.
[164] J.R. Mora, Y. Marrero-Ponce, C.R. García-Jacas, et al., Ensemble models based on QuBiLS-MAS features and shallow learning for the prediction of drug-induced liver toxicity: Improving deep learning and traditional approaches, Chem. Res. Toxicol. 33 (2020) 1855–1873.
[165] Y. Wang, X. Chen, Joint decision-making model based on consensus modeling technology for the prediction of drug-induced liver injury, J. Chem. 2021 (2021) 1–20.
[166] C.Y. Liew, Y.C. Lim, C.W. Yap, Mixed learning algorithms and features ensemble in hepatotoxicity prediction, J. Comput. Aided. Mol. Des. 25 (2011) 855–871.
[167] G. Xiong, Z. Wu, J. Yi, et al., ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucl. Acids Res. 49 (2021) W5–W14.
[168] T. Li, W. Tong, R. Roberts, et al., DeepDILI: Deep learning-powered drug-induced liver injury prediction using model-level representation, Chem. Res. Toxicol. 34 (2020) 550–565.
[169] D. Hwang, M. Jeon, J. Kang, in: Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020, pp. 323–329.
[170] H. Ma, W. An, Y. Wang, et al., Deep graph learning with property augmentation for predicting drug-induced liver injury, Chem. Res. Toxicol. 34 (2020) 495–506.
[171] T.-H. Nguyen-Vo, L. Nguyen, N. Do, et al., Predicting drug-induced liver injury using convolutional neural network and molecular fingerprint-embedded features, ACS Omega 5 (2020) 25432–25439.
[172] Z. Chen, Y. Jiang, X. Zhang, et al., ResNet18DNN: Prediction approach of drug-induced liver injury by deep neural network with ResNet18, Br. Bioinform. 23 (2022) bbab503.
[173] A. Golbamaki, E. Benfenati, A. Roncaglioni, In silico methods for carcinogenicity assessment, in: Silico Methods for Predicting Drug Toxicity, Springer, 2022, pp. 201–215.
[174] S. Limbu, S. Dakshanamurthy, Predicting environmental chemical carcinogenicity using a hybrid machine-learning approach, bioRxiv (2021) 2021.05.03.442477.
[175] Y.-W. Wang, L. Huang, S.-W. Jiang, et al., CapsCarcino: A novel sparse data deep learning tool for predicting carcinogens, Food. Chem. Toxicol. 135 (2020) 110921.
[176] T. Li, W. Tong, R. Roberts, et al., DeepCarc: Deep learning-powered carcinogenicity prediction using model-level representation, Front. Artif. Intell. (2021) 176.
[177] P. Fradkin, A. Young, L. Atanackovic, et al., A graph neural network approach for molecule carcinogenicity prediction, Bioinformatics 38 (2022) i84–i91.
[178] X. Xiang, Y. Chen, J. Gao, et al., in: Proceedings of the 2021 16th International Conference on Computer Science & Education (ICCSE), 2021, pp. 864–869.
[179] R.R. Tice, A. Bassan, A. Amberg, et al., In silico approaches in carcinogenicity hazard assessment: Current status and future needs, Comput. Toxicol. 20 (2021) 100191.
[180] F. Martin-Sanchez, K. Verspoor, Big data in medicine is driving big changes, Yearb. Med. Inform. 23 (2014) 14–20.
[181] K.P. Liao, T. Cai, G.K. Savova, et al., Development of phenotype algorithms using electronic medical records and incorporating natural language processing, BMJ 350 (2015) h1885.
[182] R.J. Carroll, A.E. Eyler, J.C. Denny, Proceedings of the AMIA Annual Symposium Proceedings, 2011, p. 189
[183] J.C. Ho, J. Ghosh, J. Sun, in: Proceedings of the Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 115–124.
[184] J.C. Ho, J. Ghosh, S.R. Steinhubl, et al., Limestone: High-throughput candidate phenotype generation via tensor factorization, J. Biomed. Inform. 52 (2014) 199–211.
[185] J. Henderson, J.C. Ho, A.N. Kho, et al., in: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), 2017, pp. 214–223.
[186] S. Gehrmann, F. Dernoncourt, Y. Li, et al., Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, PLoS One 13 (2018) e0192360.
[187] Z. Yang, M. Dehmer, O. Yli-Harja, et al., Combining deep learning with token selection for patient phenotyping from electronic health records, Sci. Rep. 10 (2020) 1–18.
[188] R. Miotto, L. Li, B.A. Kidd, et al., Deep patient: An unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 1–10.
[189] E. Guarnera, I.N. Berezovsky, Structure-based statistical mechanical model accounts for the causality and energetics of allosteric communication, PLoS Comput. Biol. 12 (2016) e1004678.
[190] B. Mieth, A. Rozier, J.A. Rodriguez, et al., DeepCOMBI: Explainable artificial intelligence for the analysis and discovery in genome-wide association studies, NAR Genom. Bioinform. 3 (2021) lqab065.
[191] E. Alsentzer, J. Murphy, W. Boag, et al., Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, Minnesota, USA, Association for Computational Linguistics, 2019, pp. 72–78.
[192] J. Listgarten, M. Weinstein, B.P. Kleinstiver, et al., Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs, Nat. Biomed. Eng. 2 (2018) 38–47.
[193] J.G. Doench, N. Fusi, M. Sullender, et al., Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9, Nat. Biotechnol. 34 (2016) 184–191.
[194] F. Boniolo, E. Dorigatti, A.J. Ohnmacht, et al., Artificial intelligence in early drug discovery enabling precision medicine, Expert Opin. Drug Discov. 16 (2021) 991–1007.
[195] A. Blasiak, J. Khong, T. Kee, CURATE, AI: Optimizing personalized medicine with artificial intelligence, SLAS Technol. 25 (2020) 95–105.
[196] A.J. Pantuck, D.K. Lee, T. Kee, et al., Modulating BET bromodomain inhibitor ZEN-3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using CURATE. AI, an artificial intelligence platform, Adv. Ther. 1 (2018) 1800104.
[197] P. Kumar, R. Benedict, F. Urzua, et al., Combination treatment significantly enhances the efficacy of antitumor therapy by preferentially targeting angiogenesis, Lab. Investig. 85 (2005) 756–767.
[198] T.M. MacDonald, B. Williams, D.J. Webb, et al., Combination therapy is superior to sequential monotherapy for the initial treatment of hypertension: A double-blind randomized controlled trial, J. Am. Heart Assoc. 6 (2017) e006986.
[199] N. Kureshi, S.S.R. Abidi, C. Blouin, A predictive model for personalized therapeutic interventions in non-small cell lung cancer, IEEE J. Biomed. Health Inform. 20 (2014) 424–431.
[200] J. Fu, A. Gucalp, M.G. Zauderer, et al., Steps in developing Watson for Oncology, a decision support system to assist physicians choosing first-line metastatic breast cancer (MBC) therapies: Improved performance with machine learning, J. Clin. Oncol. 33 (2015) 566.
[201] P. Bach, M.G. Zauderer, A. Gucalp, et al., Beyond Jeopardy!: Harnessing IBM's Watson to Improve Oncology Decision making. Editor, Book Beyond Jeopardy!:

Harnessing IBM's Watson to Improve Oncology Decision making, Series Beyond Jeopardy!: Harnessing IBM's Watson to Improve Oncology Decision Making, Am. Soc. Clin. Oncol., 2013.
[202] C. Isert, K. Atz, G. Schneider, Structure-based drug design with geometric deep learning, Curr. Opin. Struct. Biol. 79 (2023) 102548.
[203] J. Jin, D. Wang, G. Shi, et al., FFLOM: A flow-based autoregressive model for fragment-to-lead optimization, J. Med. Chem. 66 (2023) 10808–10823.
[204] H. Li, L. Zou, J.A.H. Kowah, et al., A compact review of progress and prospects of deep learning in drug discovery, J. Mol. Model 29 (2023) 117.
[205] J. Li, A. Fu, L. Zhang, An overview of scoring functions used for protein-ligand interactions in molecular docking, Interdiscip. Sci. 11 (2019) 320–328.
[206] S. Lu, X. He, D. Ni, et al., Allosteric modulator discovery: From serendipity to structure-based design, J. Med. Chem. 62 (2019) 6405–6421.
[207] D. Mucs, R.A. Bryce, The application of quantum mechanics in structure-based drug design, Expert Opin. Drug Discov. 8 (2013) 263–276.
[208] D. Ni, Z. Chai, Y. Wang, et al., Along the allostery stream: Recent advances in computational methods for allosteric drug discovery, WIREs Comput. Mol. Sci. 12 (2021) e1585.
[209] D. Ni, Y. Liu, R. Kong, et al., Computational elucidation of allosteric communication in proteins for allosteric drug design, Drug Discov. Today 27 (2022) 2226–2234.
[210] C. Pang, J. Qiao, X. Zeng, et al., Deep generative models in De Novo drug molecule generation, J. Chem. Inf. Model 64 (2024) 2174–2194.

Author profile

Jian Zhang (BRID:09778.00.20899) received his BS in pharmacology from Peking University in 2002 and PhD in medicinal chemistry from Chinese Academy of Sciences in 2007. He joined Shanghai Jiao Tong University School of Medicine as a professor in 2009 and was rewarded the Changjiang Scholars Program in 2017. He is the director of the Medicinal Chemistry & Bioinformatics Center at Shanghai Jiao Tong University. He has published 120+ peer-reviewed papers (e.g. *Nat Chem Biol, Chem*), invented 20 issued patents, and made hundreds of international invited talks. He was elected the 2017 Top-ten Science and Technology Young Scientist in China, member of the Academic Degrees Committee of the State Council in 2020, the Fellow of Royal Society of Chemistry and Excellent Research Advisor of American Chemical Society in 2022. He is the associate editor for *RSC Medicinal Chemistry*, editorial board member for *Medicinal Research Review* and other seven international journals.