CHAPTER

Deep learning in computer-aided drug design: a case study

12

Shaban Ahmad[1], Danish Aslam, Alisha Ansari, Adil Manzoor Bhat and Khalid Raza[2]
Department of Computer Science, Jamia Millia Islamia, New Delhi, India

12.1 Introduction

Computer-aided drug design (CADD) is a rapidly growing field that combines the knowledge of computational chemistry, bioinformatics, and pharmacology to aid in the discovery and development of new drugs. CADD utilizes various computational techniques and tools such as molecular modeling, docking, and machine learning (ML) to predict the properties and interactions of potential drug compounds with their biological targets (Baig et al., 2018). One of the significant advantages of CADD is its ability to significantly reduce the time and cost of drug development by allowing for the virtual screening of many compounds in silico instead of relying solely on experimental methods, which can also reduce the number of compounds that must be tested in animal models and clinical trials (Kimber et al., 2021). As a result, CADD has become an essential tool in modern pharmaceutical research (Song et al., 2009). Molecular dynamics (MD) simulations, which may foretell the motion and interactions of atoms and molecules in a biological system, and docking, which can foretell the binding of a medicinal compound to its target protein, are two examples of CADD techniques. Artificial neural networks (ANNs) and random forests are two ML techniques used in CADD for various tasks, including virtual screening, lead optimization, absorption, distribution, metabolism, and excretion (ADMET) prediction (Lavecchia, 2019; Talevi, 2018). Several CADD techniques can be utilized at different stages of the drug research and development process, including:

- Homology modeling: Homology modeling can be used to predict the structure of a protein when the structure of a related protein is known and can be used to understand the binding of drug compounds to the protein (Krieger et al., 2003).
- Virtual screening: To find compounds that might have the needed activity, a large number of compounds are computationally screened against a target protein, this process is known as virtual screening. It can be done using various techniques, such as molecular docking with various sampling algorithms, and recent trends are shown in the use of ML, which uses descriptors to learn and screen (Walters et al., 1998).
- Denovo drug design: De novo drug design is the computational design of new molecules with the desired activity against a target and can generate new candidates for drug development using various artificial intelligence (AI)-based models such as generative adversarial networks (GAN) (Popova et al., 2018).

[1]https://orcid.org/0000-0001-9832-2830.
[2]https://orcid.org/0000-0002-3646-6828.

Deep Learning Applications in Translational Bioinformatics. DOI: https://doi.org/10.1016/B978-0-443-22299-3.00012-8

- Molecular docking: Docking methods can predict the binding of a drug compound to its target protein and can be used to identify potential binding sites and evaluate the strength of the binding (Ahmad, 2022; Yadav, 2022).
- ML: In CADD, ML techniques can be used for tasks such as virtual screening, lead optimization, and ADMET prediction (Yadav, 2022).
- MD simulations: The MD simulations can predict the movement and interactions of atoms and molecules in a biological system, which can help in understanding the mechanism of drug-target interactions (DTIs) (Ahmad, Pasha Km, et al., 2022; Karwasra, Ahmad, et al., 2022; Kaul et al., 2020; Khan et al., 2021).

Overall, CADD can be used throughout the drug discovery and development process and in combination with experimental methods to aid in discovering and developing new drugs (Cardoso, 2018; Macalino et al., 2015). The basic workflow of CADD is depicted below in Fig. 12.1.

12.1.1 Drug designing approaches

Two primary methods are being used for computational drug design based on the known structure of the protein or ligand that are-

12.1.1.1 Structure-based drug design

The structure-based drug design (SBDD) uses the 3D structure of a target protein to design new compounds that bind to the protein and have the desired activity, and it uses techniques such as molecular docking or virtual screening of the compound library. SBDD is a drug discovery approach that uses the 3D structure of a biological target (usually a protein) to design small molecules that can interact with the target and modulate its activity. This approach allows scientists to design more specific and potent drugs with fewer side effects than traditional drugs

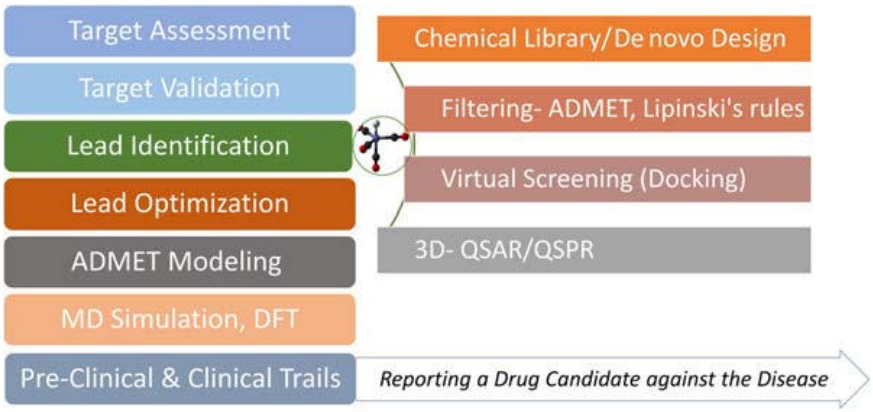

FIGURE 12.1

Showing the basic workflow of computer-aided drug design.

(Ahmad, Bano, et al., 2022; Alturki et al., 2022; Alzamami et al., 2022). The process of SBDD typically involves several steps, including:

- Identifying the target protein: The first step in SBDD is to identify a protein or macromolecule involved in a disease process that a drug could target (Yano et al., 2001).
- Determination of the 3D structure: The following stage is to ascertain the target protein's 3D structure using methods like X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy (Karwasra, Khanna, et al., 2022; Tripathi et al., 2022).
- Identifying binding sites: Once the protein structure is known, potential binding sites can be identified where small molecules could bind and modulate the protein's activity.
- Virtual screening: Computer algorithms and databases of small molecules are used to search compounds to predict to bind to the identified binding sites on the protein (Kimber et al., 2021; Walters et al., 1998).
- Optimization of leads: The most promising compounds identified from virtual screening are further optimized through iterative cycles of molecular modeling, synthesis, and testing to improve their potency, selectivity, and pharmacokinetic properties (Tripathi et al., 2022).
- Preclinical testing: Once a lead compound has been optimized, it undergoes extensive preclinical testing to evaluate its safety, efficacy, and pharmacokinetics in animal models.
- Clinical testing: If a drug candidate passes preclinical testing, it can be evaluated in clinical trials to determine its safety and efficacy in humans.

12.1.1.2 Ligand-based drug design

Ligand-based drug design (LBDD) uses the known binding properties of a ligand to a target protein to predict and design new compounds with similar properties. It uses the quantitative structure-activity relationship (QSAR) and various ML-based models to screen the compounds and predict the properties based on learning the models involving various ML techniques. Currently, the researchers are incorporating various deep learning (DL)-based techniques that learn itself and predict the best compounds. LBDD is a drug discovery approach that designs small molecules that can interact with a target protein by using information about known ligands' chemical and physical properties (molecules that bind to the target protein) (Bacilieri & Moro, 2006; Tripathi et al., 2022). LBDD does not require knowledge of the 3D structure of the target protein, making it a practical approach for identifying potential drug candidates when the protein structure is unknown or difficult to determine. The process of LBDD typically involves several steps, including:

- Identification of a known ligand: The first step in LBDD is identifying a known ligand that binds to the target protein.
- Analysis of ligand properties: The chemical and physical properties of the known ligand are analyzed to identify key features that contribute to its binding to the target protein. These properties may include the ligand's size, shape, charge distribution, and functional groups.
- Design of new ligands: Based on the analysis of the known ligand, new molecules are designed that are predicted to have similar chemical and physical properties and can bind to the target protein.
- Optimization of leads: The most promising new ligands are then further optimized through iterative cycles of molecular modeling, synthesis, and testing to improve their potency, selectivity, and pharmacokinetic properties.

- Preclinical and clinical testing: Once a lead compound has been optimized, it undergoes extensive preclinical testing to evaluate its safety, efficacy, and pharmacokinetics in animal models. If a drug candidate passes preclinical testing, it can be evaluated in clinical trials to determine its safety and efficacy in humans.

12.1.2 Why deep learning in computer-aided drug design?

DL, an emerging field of AI, is also a subfield of ML that uses ANNs with multiple layers, also known as deep neural networks (DNNs), to learn from data. These networks are designed to automatically learn complex, nonlinear relationships from substantial amounts of data, making them well-suited for tasks such as image recognition, speech recognition, and natural language processing. DL models are trained using a process called backpropagation, where the model is presented with input data and the corresponding output, and the model's parameters are adjusted to minimize the difference between the predicted output and the actual output (Bajorath, 2022; Vamathevan et al., 2019). This process is repeated with multiple data sets, allowing the model to learn and improve over time. The number of layers a network has determines its depth; the deeper a network, the more complicated problems it can answer. Due to their capacity to learn from large volumes of data, DL models are exceptionally well suited for activities like computer-aided drug creation, where a large quantity of data is needed to train the model and generate precise predictions (Dara et al., 2022). DL layers are the building blocks of a DNN. They are used to extract features from the input data and transform it into a representation that can be used to make predictions or decisions. There are several types of DL layers, including:

- Convolutional layers: These layers are used to extract features from image data. They use convolutional filters to scan the input image and extract features such as edges and textures.
- Recurrent layers: These are used to process sequential data, such as time series or natural language. They maintain a hidden state passed from one time step to the next, allowing the network to "remember" information from previous time steps.
- Fully connected layers: These layers are used to make predictions or decisions based on the features extracted by the previous layers. They are also called dense layers and are used to classify the input data.
- Pooling layers: These layers reduce the spatial dimensions of the input data, allowing the network to focus on the most prominent features.
- Normalization layers: These are used to normalize the input data and ensure that it has the same scale and distribution.
- Dropout layers: They prevent overfitting by randomly dropping out some neurons during training.

These DL layers are typically hidden in a DNN. The input data are passed through multiple layers of the network, with each layer extracting features from the input and transforming it into a new representation. The final output is produced by one or more fully connected layers that use the features extracted by the hidden layers to make predictions or decisions. The hidden layers are called so because they are not directly connected to the input or output, and the computations happening in these layers are not visible to the user (Ahmad et al., 2023). These layers are used to

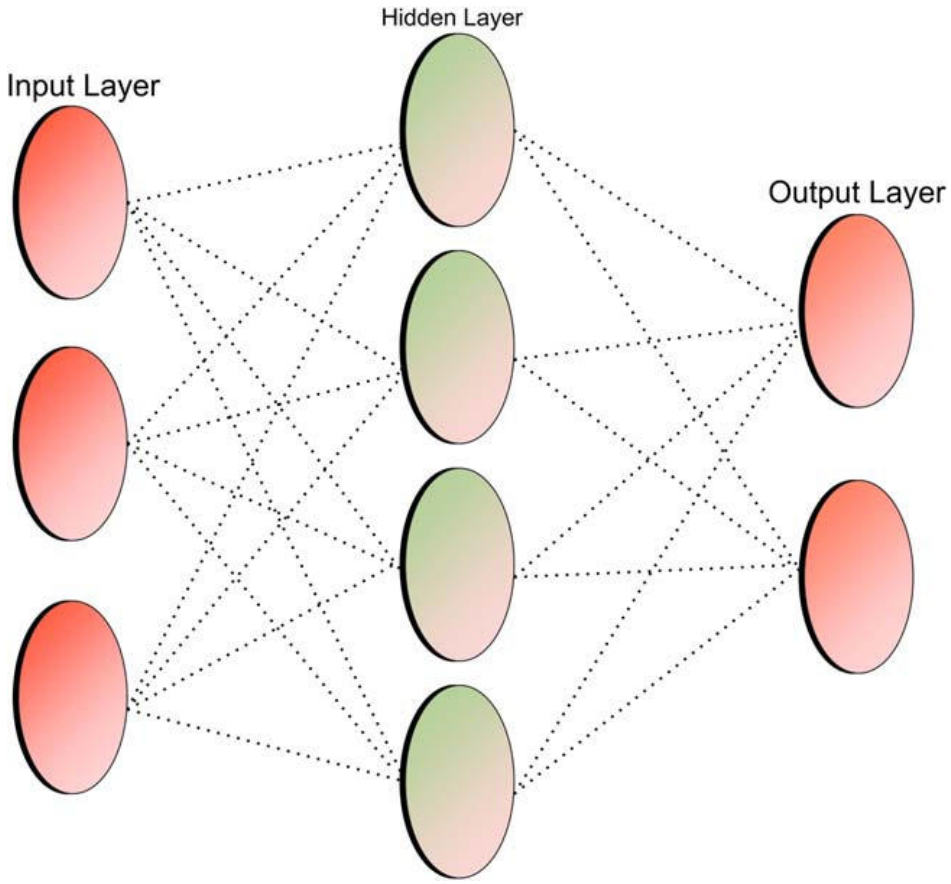

FIGURE 12.2

An illustration of a basic deep neural network architecture.

learn the abstract representation of the input data, which is then used by the final output layers to make predictions or classifications. A basic architecture of a DL model is illustrated in Fig. 12.2.

DL can significantly improve the efficiency and effectiveness of CADD. Some of the promises of DL in CADD include the following:

- Improving the accuracy of predictions: DL models can automatically learn complex, nonlinear relationships from large amounts of data, leading to more accurate predictions of the properties of drug candidates, such as binding affinity and toxicity. Several studies have demonstrated the potential of DL in improving the accuracy of predictions in CADD. In a study, a model for predicting DTIs is based on local residue patterns of involved proteins (Lee et al., 2019). In order to locally capture the residue patterns of generalized protein classes, this study developed a convolutional neural network (CNN) model that was applied to raw protein sequences with varied convolution over lengths of amino acid occurrences. Pooled convolution data were carefully examined, and the findings demonstrated that the model accurately identified protein binding sites for DTIs.

- Identifying new drug candidates: DL can analyze large libraries of compounds to identify new drug candidates with the desired properties. Several studies have demonstrated the potential of DL in identifying new drug candidates in CADD. In order to find medication candidates with the desired property outside of a data set range, a study suggested a conditional variational autoencoder (CVAE) as a generative model (Joo et al., 2020). Instead of training the model with various physical attributes for each molecule individually, this study discussed the role of molecular fingerprints and GI50 (inhibition of growth by 50%) connected to breast cancer cell lines. The CVAE model represented the desired attribute using the generated fingerprints, which were not part of the training data set.
- Reducing the time and resources required for drug discovery: DL models can analyze large amounts of data quickly and accurately, significantly reducing the time and resources required for drug discovery. Several recent studies have investigated the use of DL in CADD. In a study, a Transformer neural network architecture, a state-of-the-art approach, was suggested to handle the problem of the lengthy and costly process of drug discovery with sequence transduction tasks (Grechishnikova, 2021). This paper proposes a model that uses a self-attention strategy to capture long-range relationships between objects in succession. The model creates novel chemical structures with valuable features, perhaps requiring less time and money to develop new drugs.
- Overcoming the limitations of traditional methods: DL can overcome the limitations of traditional, rule-based methods, which can be limited by the number of known examples or the complexity of the relationships they can capture. Several studies have investigated how DL can overcome the limitations of traditional methods in CADD. DL's scope, developments, and difficulties in drug design and discovery were described in a review, with multiple experiments involving QSAR and virtual screening of repositories with thousands of compounds (Lipinski et al., 2019). This review article also discusses how DL can overcome the limitations of traditional methods in various stages of drug discovery.
- Identifying potential side effects and drug−drug interactions (DDIs): DL can be used to analyze substantial amounts of data from patient records and other sources to identify potential side effects of drugs and predict DDIs. Several studies have investigated the use of DL to predict potential side effects and DDIs in CADD. A multimodal DL framework (DDIMDL) was introduced in a study, considering various drug features combined with DL to build a model for predicting DDI-associated events (Deng et al., 2020). The model described in this article first builds several DNNs-based submodels utilizing just four types of pharmacological features: chemical substructures, targets, enzymes, and linked pathways, and then adopts a common DNN framework. In order to learn cross-modality representations of drug-drug pairs and forecast DDI events, it then integrates the submodels.
- Guiding the design of clinical trials: DL can be used to analyze data from clinical trials to guide the design of future trials and improve the chances of success. Several studies have investigated the use of DL to guide the design of clinical trials in CADD. Recent advances in AI were outlined in a review, and the key role of DL models used to reshape key steps of clinical trial design toward increasing trial success rates was discussed. It is important to note that while the studies included in the review show promise, there is still much work to be done to validate these models in real-world clinical trials and address issues such as interpretability and bias. Overall, DL has the potential to revolutionize the field of CADD by automating the process of

identifying new drug candidates, predicting their properties, and guiding the design of clinical trials, leading to the development of more effective drugs in a shorter amount of time (Krittanawong, 2022; Nagendran, n.d.).

12.2 Role of deep learning in computer-aided drug design

CADD integrates ligand-receptor interaction's biological and chemical information to unravel how well the ligand fits within the active site. The main objective of CADD is to hypothesize certain modifications in ligands in the form of chemical or functional group changes to lower energetic potential for improving the electrostatic complementarity of the ligand with the receptor. Since drug discovery and development is time-consuming, expensive, complex, and highly risky, it plunged the need and advent of CADD (Bajorath, 2022). Modern-day AI methods, especially DL, are taking up drug discovery and development space to accelerate the drug discovery and development process. DL is a subset of AI that aims at learning data precisely and provides a fruitful interpretation of data. Since a considerable amount of data is being produced by molecular biology, the DL is taking advantage of such vast volumes of data and high computational power availability to solve complexities associated with CADD. For learning model-internal representations from molecular structure, graph-based DNNs are extensively utilized in current times. DL helps to reduce the data requirements for making fruitful predictions and helps identify unknown ligand-receptor mechanisms. One of the essential components of CADD is virtual screening (Gawehn et al., 2016). DL extensively investigates a huge database of known 3D structures and evaluates them to select precise targets and accurately predict binding positions. DL can potentially predict numerous pharmacokinetic properties of drug-like molecules such as ADME. DL in the form of DNN is revolutionizing the field of proteochemometric studies. DL can learn structure representations directly without using any structure descriptors, which is employed to accelerate CADD (Anighoro, 2022). DL is making heavy inroads in retrosynthesis and reaction prediction studies. Deep CNN predicts protein-ligand interactions by scoring protein-ligand interactions. Image analysis is crucial at various stages of CADD. DL is leaving its footprints in the analysis of biological images to unravel mysteries behind protein-ligand interactions and drug mechanisms (Bai et al., 2022). DL is taking structure-based virtual screening to the next level due to the availability of high-quality models and also playing an immense role in de novo protein structure prediction, as shown in Fig. 12.3. Thus we can say that DL plays an indispensable role in the field of CADD and will continue to accelerate drug discovery, thereby reducing the time and cost involved in bringing a drug from the Lab to the bed (Barbhuiya et al., n.d.; Sabe et al., 2021).

DL has the potential to revolutionize drug discovery by accelerating the identification of promising drug candidates, reducing costs, and improving the success rate of drug development. However, developing and validating reliable DL models for drug discovery remains challenging, and careful evaluation is required to ensure the reliability of predictions made by these models. In the field of CADD, which refers to the use of computational tools and models to find, create, and optimize novel medicinal compounds, DL has many uses in CADD, some of which are covered below:

- Virtual screening: The binding affinity of small compounds to protein targets can be predicted using DL, and it can be accomplished by using a large dataset of known DTIs to build a DL

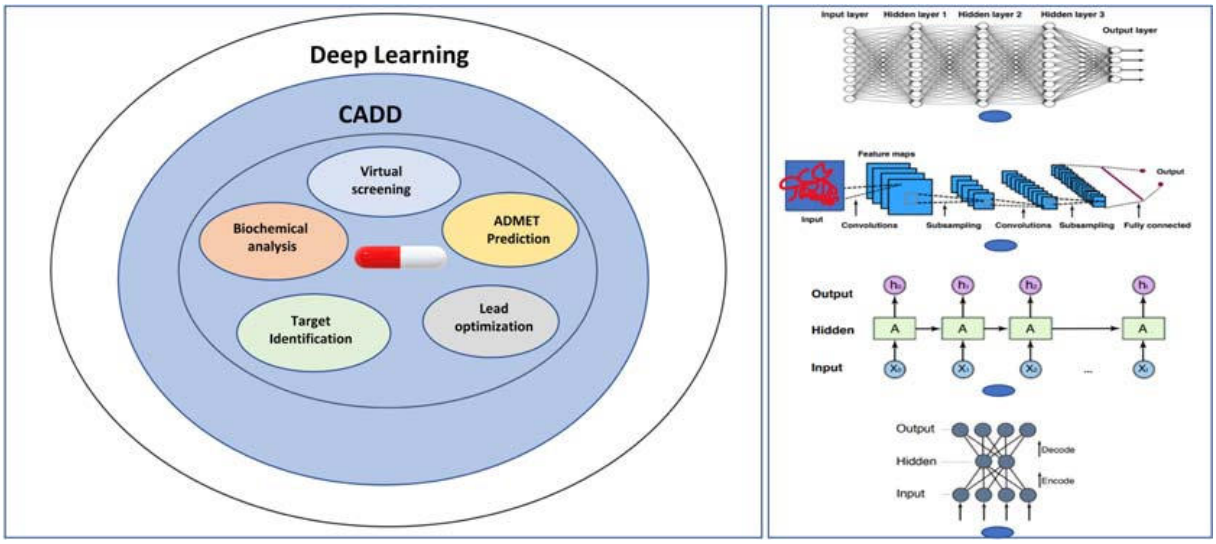

FIGURE 12.3

The left hand of this figure displays the role of DL at various stages of CADD, and the right-hand side displays different DL algorithms like DNN, CNN, RNN, and autoencoder. *CADD*, Computer-aided drug design; *CNN*, convolutional neural network; *DL*, deep learning; *DNNs*, deep neural networks.

model. The trained model can then determine the predicted binding affinity of novel compounds to target proteins. In order to find potential drug candidates for further development, this can be a helpful technique.

- Molecular generation: New pharmacological compounds tailored for particular target proteins can be created using DL, and it can be accomplished by using a massive dataset of well-known pharmacological compounds and their related properties to train a DL model. The novel compounds can then be created using the learned model and tailored for particular target proteins.
- SBDD: DL can forecast protein targets' 3D structures. A DL model can be trained on a sizable dataset of well-known protein structures to accomplish this. The structure of new protein targets can be predicted using the trained model. This can be a valuable technique for creating medications tailored for particular protein targets.
- ADME prediction: The ADMET characteristics of pharmacological compounds can also be predicted using DL. A DL model can be trained using a sizable dataset of known drug compounds and the related ADME features. The ADME properties of fresh drug candidates can then be predicted using the learned model.
- Multitask learning: DL is used to predict several drug molecules, including binding affinity, solubility, and toxicity, and it can be accomplished by using a massive dataset of pharmacological compounds and their related attributes to build a DL network. The trained model can then be used to predict several features of novel drug candidates simultaneously.
- Transfer learning: DL can transfer knowledge from one drug discovery task to another. For example, a DL model trained on one protein target can be fine-tuned to predict the binding

affinity of another protein target. This can be a valuable tool in drug discovery, mainly when data are limited for a particular target.

- Drug repurposing: DL can be utilized to find novel therapeutic applications for already available medications, and it can be accomplished by using a huge dataset of pharmacological compounds and their related attributes to build a DL network that can be utilized to find novel therapeutic applications for already available medications. It can be accomplished using a massive dataset of pharmacological compounds and their related attributes to build a DL network. The trained model can then predict the therapeutic uses of existing drugs not initially developed for that purpose.
- Interpretability: DL models can provide insights into the mechanism of action of drug molecules. For example, DL models can identify the key features of drug molecules responsible for their binding affinity to target proteins. It can provide insights into the underlying biology of the disease and inform drug discovery efforts.

Overall, by accurately predicting important characteristics of therapeutic compounds and protein targets, DL has the potential to speed up the drug discovery process dramatically. However, it is crucial to remember that DL models are only as good as the data they are trained on. Therefore, proper validation is necessary to guarantee the accuracy of DL model predictions in CADD.

12.3 Software tools, web servers, and package

DL is a rapidly growing field that has led to significant advancements in various applications, including drug discovery. A key aspect of DL is the availability of powerful tools, web servers, and packages that enable researchers and developers to implement and deploy DL models quickly. This survey aims to provide an overview of the current state-of-the-art DL tools, web servers, and packages (Barbhuiya et al., 2022). The focus is also on the most popular and widely used tools and platforms that comprehensively evaluate their language and models. The survey will cover various topics, including DL tools, frameworks, and web servers for deploying DL models and packages for various applications. The goal is to provide a comprehensive resource for researchers, developers, and practitioners in the DL field and help them make informed decisions when choosing the right tools and platforms for their projects. It is important to note that drug design and discovery are rapidly evolving, and new tools and web servers are constantly being developed. Moreover, the list of DLBDD tools, web servers, and software packages discussed in this section is the result of a survey and may contain results based on some bias related to the discussed parameters.

12.3.1 Parameters

Several parameters were considered to evaluate these tools, web servers, and software packages-

Algorithm: Different tools and web servers use different algorithms to predict the binding affinity of small molecules for a protein target or to predict the activity of small molecules against different targets. It is essential to understand the algorithm used in a tool or webserver and whether it suits the specific use case.

- Language: The tools, web servers, and software packages are implemented in different programming languages.
- Database: Some web servers provide access to large databases of small molecules, which can be helpful for virtual screening and data size and quality matters.
- Performance: Different tools, web servers, and software packages have different performance characteristics. When evaluating a tool or web server, it is essential to consider factors such as accuracy, speed, and scalability.
- User interface: The user interface of a tool or webserver can affect its ease of use and accessibility. It is essential to consider whether the user interface is intuitive and easy to use when evaluating a tool or web server.
- Cost: Some tools, web servers, and software packages are open-source and freely available, while others are commercial and require a license. It is essential to consider a tool's or web server's cost when evaluating it.
- Support: Different tools, web servers, and software packages have different levels of support available. When evaluating it, it is essential to consider whether good documentation and support are available for a tool or web server.

12.3.2 Tools

The tools considered in this survey are listed in Table 12.1. This list is not exhaustive; new tools and web servers may be developed to suit the specific use case better.

Table 12.1 Tools using deep learning models for processes included in drug discovery.

S. no.	Name	Description	References
1	DeepDTA	A DL-based tool for predicting drug-target interactions	Öztürk et al. (2018)
2	MolDQN	Reinforcement learning-based tool for virtual screening of small molecules	Zhenpeng et al. (2019)
3	DeepChem	An open-source toolkit for DL in drug discovery and materials science	Ramsundar (2016, 2019)
4	MolNetEnhancer	A DL-based tool for enhancing molecular representations	Ernst et al. (2019)
5	DeepSite	A DL-based tool for predicting protein-ligand binding sites	Zeng et al. (2020)
6	Pafnucy	DL-based tool for predicting protein-ligand binding affinity	Trott and Olson (2010)
7	GENTRL	A DL-based tool for drug discovery and design	Zhavoronkov et al. (2019)
8	MolCycleGAN	A DL-based tool for generating new small molecules using a GAN model	Maziarka et al. (2020)
9	Deep Docking	DL platform for Augmentation of SBDD	Gentile et al. (2020)
10	DeepDrug3D	A DL-based tool for drug design and discovery	Pu et al. (2019)

SBDD, *Structure-based drug design.*

Table 12.2 Software Packages implementing deep learning models for processes in drug discovery.

S. no.	Name	Description	Sources
1	RDKit	Python-based Cheminformatics and ML tools	https://www.rdkit.org/
2	Open Babel	A C++ chemical toolbox designed to speak the many languages of chemical data	http://openbabel.org/
3	Keras	A Python-based high-level neural networks API	https://keras.io/
4	Theano	Python library that allows defining optimizing, and evaluating mathematical expressions	http://deeplearning.net/software/theano/
5	MATLAB	Interactive environment for numerical computation and visualization	https://www.mathworks.com/products/matlab.html

12.3.3 Packages

A list of some libraries and software packages for DL-based drug design, along with their associated language, algorithm, and references, has been listed in Table 12.2. This is not an exhaustive list; new packages may be developed to suit the specific use case better.

12.4 Promises and challenges of deep learning in computer-aided drug design

DL, a subset of ML that utilizes ANNs with multiple layers, has the potential to revolutionize the field of CADD by harnessing the power of big data and complex models. The ability of DL to learn intricate patterns and features in large compounds and experimental data datasets can potentially improve the efficiency and effectiveness of CADD. Although DL is currently revolutionizing the field of CADD with greater accuracy, DL in CADD is promising in several ways:

- Improved accuracy: DL can improve virtual screening accuracy by identifying patterns and features in large datasets of compounds that are not easily captured by traditional methods. This can lead to the identification of new and more potent lead compounds.
- Lead optimization: DL can be used to optimize the properties of lead compounds, such as binding affinity, selectivity, and pharmacokinetics, by using large datasets obtained from experimental data and computational simulations. This can aid in improving the efficacy and lowering the toxicity of lead compounds.
- Virtual screening: DL can also improve the accuracy of virtual screening by identifying patterns and features in large datasets of compounds that are not easily captured by traditional methods. This can lead to the identification of new and more potent compounds.
- ADMET prediction: DL can be used to predict lead compounds' potential toxicity and pharmacokinetic properties by analyzing large datasets of experimental data and using them to train models that can predict the ADMET properties of new compounds. The early prediction of ADMET properties via DL enables the identification of potential risks that can hamper the drug

discovery process in later stages. Thus, it can help ensure safety, enhance accuracy, and reduce costly failures in later stages.
- Handling large and complex data: DL algorithms can handle large and complex datasets; thus, they can be used to analyze diverse types of data, such as genomics, proteomics, metabolomics, and other omics, which can provide valuable insights into drug development.
- Target identification: DL can identify potential drug targets by analyzing large datasets of biological data, such as gene expression data, and using them to train DL models. This can help to identify new targets and develop new drugs for diseases that have previously been difficult to treat.

The revolution comes with challenges, so DL is facing a massive challenge for now. However, it facilitates more features with robust datasets. Here are a few challenges listed-

- Data availability: DL models require a large amount of data to train the models. Although the chemical space of possible drug-like molecules is $\sim 10^{60}$, there are currently 13,117 FDA-approved human drugs only because CADD is costly and time-consuming, with an average cost of $2.6 billion for designing one drug in 10 years. There is a shortage of well-characterized drug-like molecules upon which the DL models can be trained.
- Computational cost: DL models require high-end computational resources to train on, especially for drug discovery-like tasks where models are large generally because of the high number of features, for example, molecular descriptors. The specific requirements depend on the model's size and complexity and the amount of data used for training.
- Data quality: The data used for training DL models should be high quality, accurate, and representative of the real-world scenario. However, in drug discovery, data could be noisy, biased, or incomplete, leading to poor performance of the model and inaccurate predictions.
- Data heterogeneity: Drug discovery data can be highly heterogeneous, coming from various sources such as images, graphs, time-series data (e.g., SMILES strings), and many more making it challenging to preprocess and integrate the data for DL models.
- Model fitting: DL models trained on a specific dataset may not generalize well to test datasets, the problem, also known as overfitting, is quite common. For drug discovery, there is a shortage of well-characterized molecules specific to research/target proteins, and it is challenging to build a model with acceptable accuracy on the train and test data.
- Model interpretation: Understanding how a DL model makes its predictions is crucial for drug discovery, where the safety and efficacy of drug candidates are critical. However, DL models are often considered a "black box" and can be challenging to interpret. For example, DL models in drug discovery like CNN, recurrent neural network (RNN), and graph neural network (GNN) can perform feature selection automatically instead of using manually drawn features of molecules.
- Safety and ethical concerns: The use of DL in drug discovery raises important ethical and safety issues, such as ensuring that the models do not perpetuate existing biases, resulting in inaccurate predictions of potential drug candidates. However, such biases can be controlled by wet-lab experimental validation using in vitro and in vivo techniques.

Thus we can say that DL can change how we think of drug design, discovery, and development shortly by its highly robust, precise, and accurate nature. However, at the same time, we must address the challenges DL poses in drug discovery and development to harness the potential benefits.

12.5 Case studies

In this chapter, we kept four case studies to understand the concept of DL in computer-aided drug designing better.

12.5.1 Case study #1: "GNNEXPLAINER: an accurate and effective method for explaining graph neural networks"

GNNs have shown great success in various graph-based tasks, but their decision-making process is often opaque and difficult to interpret (Jiménez-Luna et al., 2020). This lack of interpretability raises concerns about the trustworthiness and fairness of the models. A recent study presented GNNEXPLAINER, a method for explaining GNN predictions for node and graph classification tasks. The authors compare their method against two alternative baseline approaches: GRAD, a gradient-based method that computes the gradient of the GNN's loss function concerning the adjacency matrix and the associated node features, and ATT, a graph attention GNN that learns attention weights for edges in the computation graph. Another notable contribution of this study is the use of real-world graph classification datasets, MUTAG and REDDIT-BINARY, which are commonly used in the literature to evaluate the performance of GNNs. The results demonstrate that GNNEXPLAINER can provide meaningful explanations for the predictions made by GNNs on these datasets, which can be helpful for domain experts in understanding the underlying patterns and relationships in the data. Overall, the study addresses a significant challenge in explainable AI by proposing a method for generating explanations for predictions made by GNNs on graph-structured data. The authors thoroughly evaluate their method on both synthetic and real-world datasets, and the results demonstrate its effectiveness and accuracy in identifying important features and structures in the graphs. The qualitative and quantitative analysis of the explanations generated by GNNEXPLAINER demonstrates its superiority over the baseline approaches.

12.5.2 Case study #2: "CADD techniques and applications in Alzheimer's drug discovery"

The use of DL in drug discovery has become increasingly popular in recent years, thanks to the availability of powerful computing tools such as GPUs (Dorahy et al., 2023). DL involves using multilayered neural networks to analyze large, multidimensional datasets. In drug design, CNNs are commonly used to extract features from molecular graphs and predict pharmacokinetic properties or binding affinities of ligand-protein complexes. DL has shown promising results in predicting blood-brain barrier permeability, and its applications extend to image recognition and natural language processing. Several CADD studies have used various drug targets, such as acetylcholinesterase, serotonin transporter, beta-secretase 1, and glycogen synthase kinase-3, to create multitarget activity models for new therapeutics. These models have shown promise in identifying ligands with drug properties and favorable efficacy for further screening. A multitarget activity 3D-QSAR model was used in a study to identify potential therapeutics for Alzheimer's disease. The model was built using IC50 data from ChEMBL, and both multilinear regression and ANN models were used. Ligands were converted to 3D structures using OpenBabel, and molecular descriptors were generated with FQSARModel. After validation, ANN models performed better than multilinear regression models

in virtual screening against over 20,000 compounds from the ZINC database docked against all four proteins (AChE, SERT BACE1, and GSK3β). Fifty-seven compounds with drug properties and favorable ligand efficacy were screened against QSAR models, identifying five promising ligands that could target at least three of the four proteins involved in Alzheimer's disease pathology. One compound called ZINC4027357 demonstrated inhibition of AChE and BACE1; none had inhibitory properties against SERT or GSK3β within selected potencies. CADD is helping to hasten the process of finding new lead molecules to become approved drugs, and newer techniques, such as DL, are expected to increase the efficiency with which studies can be carried out.

12.5.3 Case study #3: "Identification of structural motifs in dual-target compounds using explainable machine learning"

Dual-target compounds have gained increasing attention in drug discovery due to their potential to treat complex diseases with multiple pathological targets (Feldmann et al., 2021). However, identifying these compounds is challenging due to their complex mechanisms of action and multi-DTIs. In a study, explainable ML methods were applied to predict dual-target compounds and explore the structural features that contribute to their activity. The data sets comprising DT-CPDs (dual-target compounds) and corresponding ST-CPDs (single-target compounds) with activity against target pairs 1 (MAOB and A2aR) and 2 (MAOB and AChE) were discerned with the use of balanced random forest (BRF) classification models to distinguish. The choice of target pairs was based on specific criteria, comprising sharing a target (MAOB), belonging to different protein classes, and being implicated in diseases falling into the same medicinal area. Using BRFs allowed for calculating accurate local SVs (Shapley value) for a model explanation. SV analysis quantifies the assistance of different features (e.g., layered atom environments) to a compound's prediction, encompassing the absence of certain features, which can help recognize structural motifs characteristic of DT-CPDs. In conclusion, the study contributed an in-depth investigation to explore the structural motifs signatures of compounds with well-defined activity against more than one target. Highly accurate ML prediction models were derived for the coinciding target pairs system, and the SV concept was used for model explanation and extended for global feature analysis. The study identified molecular representation features that determine ML-based predictions of DT- versus ST-CPDs and quantified the influence of these features in present and absent test compounds. Small numbers of features whose presence in DT-CPDs and absence in corresponding ST-CPDs were definitive for accurate predictions, thus providing apparent grounds for DT activity. These features were specific for overlapping yet unique DT activities, forming coherent substructures in DT-CPDs. Two structural motifs, caffeine and coumarin fragments, occurred as largely determined accurate predictions of DT-CPDs and represented characteristic substructures conferring different DT activities. The reproducible analysis scheme applies to other target combinations and compound features. Further, the same is shown in Fig. 12.4.

12.5.4 Case study #4: "Drug design by machine learning: support vector machines for pharmaceutical data analysis"

AI and ML methods in the medical field have become increasingly popular in recent years (Staszak et al., 2022). These methods have been applied in drug discovery, lead optimization and synthesis,

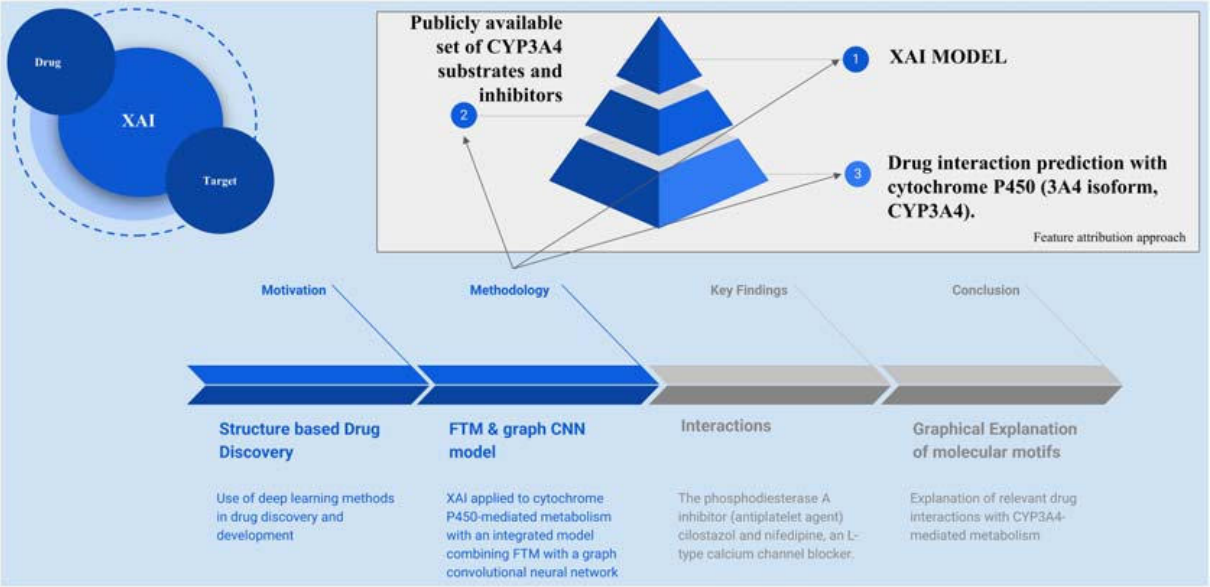

FIGURE 12.4

Showing the graphical abstract for the case study of explainable machine learning predictions of dual-target compounds reveal characteristic structural features.

medical image analysis, diabetic diseases, and oncology research. The pharmaceutical industry has shown particular interest in ML due to the potential for identifying reliable therapeutic inferences that can lead to the development of appropriate drugs. The emergence of high-performance research approaches in biology and diseases and increased computing capabilities have created chances to build complex AI systems based on large data sets, including text, images, biometric data, and multidimensional elements. A study presented the use of support vector machines (SVMs) to predict the activity of molecules against a specific target, such as an enzyme or receptor, based on their chemical structure and other properties. The advantages of SVMs, including their ability to handle high-dimensional data, their flexibility in handling different types of data, and their ability to handle nonlinear relationships between variables, were highlighted. As a proof-of-concept, a comparison was made based on the performance of three different ML methods in predicting dihydrofolate reductase (DHFR) inhibition by pyrimidines: SVM, RBF kernel, and C5.0 decision tree. The SVM method was the most effective deterministic learning algorithm, producing reproducible results with the lowest model error and shortest calculation time compared to the other two methods. Furthermore, the study suggested that this methodology could potentially be used to predict the properties of drugs in terms of their toxicity. It is worth noting that additional validation studies would be necessary to confirm these findings' generalizability and assess the performance of these methods on other datasets. Overall, the study provided some promising results regarding the application of SVMs in drug toxicity prediction, but further research is needed to confirm and expand upon these findings. Further, the same is shown in Fig. 12.5.

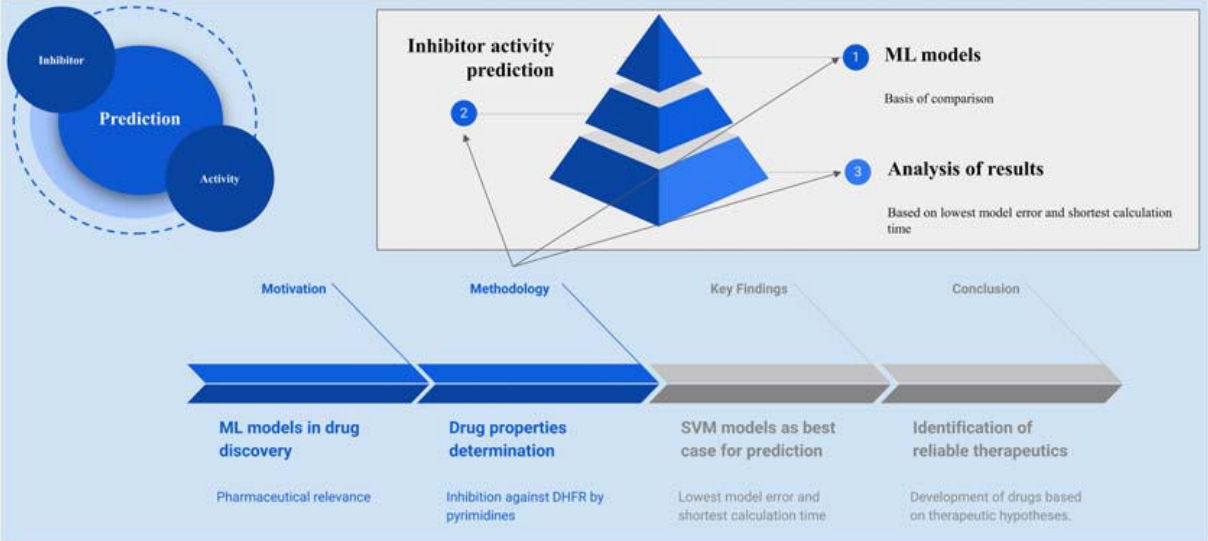

FIGURE 12.5

Showing the graphical abstract for the case study of machine learning in drug design: use of artificial intelligence to explore the chemical structure–biological activity relationship.

12.6 Conclusion

Deep learning (DL) is completely revolutionizing the CADD with the advancement of various structural learning, from finding new druggable targets to screening the drug candidates and showcasing the best among the library of millions of compounds. It also has a vast potential to minimize the cost with accuracy for the drug design, and it has been incorporated into clinical trials to make the process smooth and accurate by removing accessibility errors. After so much advancement, there is much more to work to make the DL proof, as the traditional methods of many CADD subareas are still questionable.

References

Ahmad, N., Arya, S., & Singh, D. (2023). *Predicting risky environment for child inside house using deep learning 2023 international conference on emerging smart computing and informatics (ESCI)*. IEEE.

Ahmad, S. (2022). Molecular dynamics simulation and docking analysis of NF-κB protein binding with sulindac acid. *Bioinformation*, *18*(3), 170–179. Available from https://doi.org/10.6026/97320630018170.

Ahmad, S., Bano, N., Qazi, S., Yadav, M.K., Ahmad, N., Raza, K. (2022). Multitargeted molecular dynamic understanding of butoxypheser against SARS-CoV-2: An in silico study. Natural Product Communications. 17 (7), 1934578X2211154. Available from https://doi.org/10.1177/1934578x221115499.

Ahmad, S., Pasha Km, M., Raza, K., Rafeeq, M. M., Habib, A. H., Eswaran, M., & Yadav, M. K. (2022). Reporting dinaciclib and theodrenaline as a multitargeted inhibitor against SARS-CoV-2: An in-silico

study. *Journal of Biomolecular Structure and Dynamics*, 1–11. Available from https://doi.org/10.1080/07391102.2022.2060308.

Alturki, N. A., Mashraqi, M. M., Alzamami, A., Alghamdi, Y. S., Alharthi, A. A., Asiri, S. A., Ahmad, S., & Alshamrani, S. (2022). In-silico screening and molecular dynamics simulation of drug bank experimental compounds against SARS-CoV-2. *Molecules (Basel, Switzerland)*, *27*(14), 4391. Available from https://doi.org/10.3390/molecules27144391.

Alzamami, A., Alturki, N. A., Alghamdi, Y. S., Ahmad, S., Alshamrani, S., Asiri, S. A., & Mashraqi, M. M. (2022). Hemi-babim and fenoterol as potential inhibitors of MPro and papain-like protease against SARS-CoV-2: An in-silico study. *Medicina*, *58*(4), 515. Available from https://doi.org/10.3390/medicina58040515.

Anighoro, A. (2022). *Deep learning in structure-based drug design methods in molecular biology* (pp. 261–271). Humana Press Inc. Available from http://www.springer.com/series/7651, 10.1007/978-1-0716-1787-8_11.

Bacilieri, M., & Moro, S. (2006). Ligand-based drug design methodologies in drug discovery process: An overview. *Current Drug Discovery Technologies*, *3*(3), 155–165. Available from https://doi.org/10.2174/157016306780136781Italy, http://docstore.ingenta.com/cgi-bin/ds_deliver/1/u/d/ISIS/36105468.1/ben/cddt/2006/00000003/00000003/art00001/2B915E658EF62D8A1173853660FC8A8872600B95D9.pdf?link = htt://http://www.ingentaconnect.com/error/delivery&format = pdf.

Baig, M. H., Ahmad, K., Rabbani, G., Danishuddin, M., & Choi, I. (2018). Computer aided drug design and its application to the development of potential drugs for neurodegenerative disorders. *Current Neuropharmacology*, *16*(6), 740–748. Available from https://doi.org/10.2174/1570159X15666171016163510, http://www.eurekaselect.com/603/journal/current-neuropharmacology.

Bai, Q., Liu, S., Tian, Y., Xu, T., Banegas-Luna, A. J., Pérez-Sánchez, H., Huang, J., Liu, H., & Yao, X. (2022). Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(3). Available from https://doi.org/10.1002/wcms.1581, http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1759-0884.

Bajorath, J. (2022). Deep machine learning for computer-aided drug design. *Frontiers in Drug Discovery*, *2*. Available from https://doi.org/10.3389/fddsv.2022.829043.

Barbhuiya, R.K., & Ahmad, N. (2021). IoT applications in translational bioinformatics. In *Translational bioinformatics in healthcare and medicine*. Academic Press.

Barbhuiya, R. K., Ahmad, N., & Akram, W. (2022). *Application of convolutional neural networks in cancer diagnosis. Computational intelligence in oncology: Applications in diagnosis, prognosis and therapeutics of cancers* (pp. 95–109). Singapore: Springer.

Cardoso, M.J. (2018). *Computational pathology and ophthalmic medical image analysis: first international workshop, COMPAY 2018, and 5th international workshop, OMIA 2018, Held in Conjunction with MICCAI*. Springer.

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. *Artificial Intelligence Review*, *55*(3), 1947–1999. Available from https://doi.org/10.1007/s10462-021-10058-4, http://www.springer.com/journal/10462.

Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., & Liu, S. (2020). A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics (Oxford, England)*, *36*(15), 4316–4322. Available from https://doi.org/10.1093/bioinformatics/btaa501, http://bioinformatics.oxfordjournals.org/.

Dorahy, G., Chen, J. Z., & Balle, T. (2023). Computer-aided drug design towards new psychotropic and neurological drugs. *Molecules (Basel, Switzerland)*, *28*(3). Available from https://doi.org/10.3390/molecules28031324, http://www.mdpi.com/journal/molecules.

Ernst, M., Kang, K. B., Caraballo-Rodríguez, A. M., Nothias, L.-F., Wandy, J., Chen, C., Wang, M., Rogers, S., Medema, M. H., Dorrestein, P. C., & van der Hooft, J. J. J. (2019). MolNetEnhancer: Enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites*, *9*(7), 144. Available from https://doi.org/10.3390/metabo9070144.

Feldmann, C., Philipps, M., & Bajorath, J. (2021). Explainable machine learning predictions of dual-target compounds reveal characteristic structural features. *Scientific Reports*, *11*(1). Available from https://doi.org/10.1038/s41598-021-01099-4, http://www.nature.com/srep/index.html.

Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. *Molecular Informatics*, *35*(1), 3−14. Available from https://doi.org/10.1002/minf.201501008, http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1868-1751.

Gentile, F., Agrawal, V., Hsing, M., Ton, A. T., Ban, F., Norinder, U., Gleave, M. E., & Cherkasov, A. (2020). Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, *6*(6), 939−949. Available from https://doi.org/10.1021/acscentsci.0c00229, http://pubs.acs.org/journal/acscii.

Grechishnikova, D. (2021). Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific Reports*, *11*(1). Available from https://doi.org/10.1038/s41598-020-79682-4, http://www.nature.com/srep/index.html.

Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, *2*(10), 573−584. Available from https://doi.org/10.1038/s42256-020-00236-4, https://www.nature.com/natmachintell/.

Joo, S., Kim, M. S., Yang, J., & Park, J. (2020). Generative model for proposing drug candidates satisfying anticancer properties using a conditional variational autoencoder. *ACS Omega*, *5*(30), 18642−18650. Available from https://doi.org/10.1021/acsomega.0c01149, http://pubs.acs.org/journal/acsodf.

Karwasra, R., Ahmad, S., Bano, N., Qazi, S., Raza, K., Singh, S., & Varma, S. (2022). Macrophage-targeted punicalagin nanoengineering to alleviate methotrexate-induced neutropenia: A molecular docking, DFT, and MD simulation analysis. *Molecules (Basel, Switzerland)*, *27*(18), 6034. Available from https://doi.org/10.3390/molecules27186034.

Karwasra, R., Khanna, K., Singh, S., Ahmad, S., & Verma, S. (2022). *The incipient role of computational intelligence in oncology: Drug designing, discovery, and development* (pp. 369−384). Springer Science and Business Media LLC. Available from 10.1007/978-981-16-9221-5_21.

Kaul, T., Eswaran, M., Ahmad, S., Thangaraj, A., Jain, R., Kaul, R., Raman, N. M., & Bharti, J. (2020). Probing the effect of a plus 1bp frameshift mutation in protein-DNA interface of domestication gene, NAMB1, in wheat. *Journal of Biomolecular Structure and Dynamics*, *38*(12), 3633−3647. Available from https://doi.org/10.1080/07391102.2019.1680435, http://www.tandfonline.com/loi/tbsd20.

Khan, F. N., Khanam, A. A., Ramlal, A., & Ahmad, S. (2021). *A review on predictive systems and data models for COVID-19 studies in computational intelligence* (pp. 123−164). Springer Science and Business Media Deutschland GmbH. Available from http://www.springer.com/series/7092, 10.1007/978-981-15-8534-0_7.

Kimber, T. B., Chen, Y., & Volkamer, A. (2021). Deep learning in virtual screening: Recent applications and developments. *International Journal of Molecular Sciences*, *22*(9). Available from https://doi.org/10.3390/ijms22094435, https://www.mdpi.com/1422-0067/22/9/4435/pdf.

Krieger, E., Nabuurs, S. B., & Vriend, G. (2003). Homology modeling. *Methods of Biochemical Analysis*, *44*, 509−523. Available from https://doi.org/10.1002/0471721204.ch25.

Krittanawong, C. (2022). The next step in deep learning-guided clinical trials. *Nature Cardiovascular Research*, *1*(4), 286−288. Available from https://doi.org/10.1038/s44161-022-00044-6.

Lavecchia, A. (2019). Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discovery Today*, *24*(10), 2017−2032. Available from https://doi.org/10.1016/j.drudis.2019.07.006, http://www.elsevier.com/locate/drugdiscov.

Lee, I., Keum, J., Nam, H., & Briggs, J. M. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, *15*(6), e1007129. Available from https://doi.org/10.1371/journal.pcbi.1007129.

Lipinski, C. F., Maltarollo, V. G., Oliveira, P. R., da Silva, A. B. F., & Honorio, K. M. (2019). Advances and perspectives in applying deep learning for drug design and discovery. *Frontiers in Robotics and AI* (6). Available from https://doi.org/10.3389/frobt.2019.00108, http://journal.frontiersin.org/journal/robotics-and-ai.

Macalino, S. J. Y., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, *38*(9), 1686−1701. Available from https://doi.org/10.1007/s12272-015-0640-5, http://www.springerlink.com/content/0253-6269.

Maziarka, Ł., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., & Warchoł, M. (2020). Mol-CycleGAN: A generative model for molecular optimization. *Journal of Cheminformatics*, *12*(1). Available from https://doi.org/10.1186/s13321-019-0404-1, https://jcheminf.biomedcentral.com/.

Nagendran, M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ (Clinical Research ed.)*. Available from https://doi.org/10.1136/bmj.m689.

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug-target binding affinity prediction. *Bioinformatics*, *34*(17), i821−i829. Available from http://bioinformatics.oxfordjournals.org/, Oxford University Press Turkey 9 1 2018/09/01, 10.1093/bioinformatics/bty593.

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances*, *4*(7). Available from https://doi.org/10.1126/sciadv.aap7885, http://advances.sciencemag.org/content/4/7/eaap7885/tab-pdf.

Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H.-C., Brylinski, M., & Schneidman-Duhovny, D. (2019). DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Computational Biology*, *15*(2), e1006718. Available from https://doi.org/10.1371/journal.pcbi.1006718.

Ramsundar, B. (2016). *Democratising deep-learning for drug discovery, quantum chemistry, materials science and biology GitHub repository*.

Ramsundar, B. (2019). *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*.

Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., & Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European Journal of Medicinal Chemistry*, *224*, 113705. Available from https://doi.org/10.1016/j.ejmech.2021.113705.

Song, C. M., Lim, S. J., & Tong, J. C. (2009). Recent advances in computer-aided drug design. *Briefings in Bioinformatics*, *10*(5), 579−591. Available from https://doi.org/10.1093/bib/bbp023.

Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., & Tylkowski, B. (2022). Machine learning in drug design: Use of artificial intelligence to explore the chemical structure−biological activity relationship. *Wiley Interdisciplinary Reviews: Computational Molecular Science.*, *12*(2). Available from https://doi.org/10.1002/wcms.1568, http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1759-0884.

Talevi, A. (2018). *Computer-aided drug design: An overview methods in molecular biology* (pp. 1−19). Humana Press Inc. Available from http://www.springer.com/series/7651, 10.1007/978-1-4939-7756-7_1.

Tripathi, M. K., Ahmad, S., Tyagi, R., Dahiya, V., & Yadav, M. K. (2022). *Fundamentals of molecular modeling in drug design Computer Aided Drug Design (CADD): From ligand-based methods to structure-based approaches* (pp. 125−155). Elsevier. Available from https://www.sciencedirect.com/book/9780323906081, 10.1016/B978-0-323-90608-1.00001-0.

Trott, O., & Olson, A. J. (2010). Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), 455−461. Available from http://www3.interscience.wiley.com/cgi-bin/fulltext/122439542/PDFSTART.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development.

Nature Reviews. Drug Discovery, *18*(6), 463−477. Available from https://doi.org/10.1038/s41573-019-0024-5, http://www.nature.com/nrd/index.html.

Walters, W. P., Stahl, M. T., & Murcko, M. A. (1998). Virtual screening—An overview. *Drug Discovery Today*, *3*(4), 160−178. Available from https://doi.org/10.1016/s1359-6446(97)01163-x.

Yadav, M. K. (2022). Predictive modeling and therapeutic repurposing of natural compounds against the receptor-binding domain of SARS-CoV-2. *Journal of Biomolecular Structure and Dynamics*, 1−13. Available from https://doi.org/10.1080/07391102.2021.2021993.

Yano, H., Wong, J. H., Lee, Y. M., Cho, M.-J., & Buchanan, B. B. (2001). A strategy for the identification of proteins targeted by thioredoxin. *Proceedings of the National Academy of Sciences*, *98*(8), 4794−4799. Available from https://doi.org/10.1073/pnas.071041998.

Zeng, W., Wang, Y., & Jiang, R. (2020). Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics (Oxford, England)*, *36*(2), 496−503. Available from https://doi.org/10.1093/bioinformatics/btz562, http://bioinformatics.oxfordjournals.org/.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., … Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, *37*(9), 1038−1040. Available from https://doi.org/10.1038/s41587-019-0224-x, http://www.nature.com/nbt/index.html.

Zhenpeng, Z., Steven, K., Li, L., Richard, Z., & Patrick, R. (2019). Optimisation of molecules via deep reinforcement learning. *Scientific Reports*, *9*, 1−10. Available from https://doi.org/10.1038/s41598-019-47148-x.