Methods 219 (2023) 82–94

Contents lists available at ScienceDirect

Methods

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

Exploring the artificial intelligence and machine learning models in the context of drug design difficulties and future potential for the pharmaceutical sectors

Periyasamy Natarajan Shiammala [a], Navaneetha Krishna Bose Duraimutharasan [a], Baskaralingam Vaseeharan [b], Abdulaziz S. Alothaim [c], Esam S. Al-Malki [c], Babu Snekaa [d], Sher Zaman Safi [e], Sanjeev Kumar Singh [f], Devadasan Velmurugan [g], Chandrabose Selvaraj [d,h,*]

[a] Department of Information Technology, AMET Deemed to be University, Kanathur, Chennai, Tamil Nadu 603112, India
[b] Department of Animal Health and Management, Science Block, Alagappa University, Karaikudi, Tamil Nadu 630 003, India
[c] Department of Biology, College of Science in Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia
[d] Laboratory for Artificial Intelligence and Molecular Modelling, Department of Pharmacology, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamil Nadu 600077, India
[e] Faculty of Medicine, Bioscience and Nursing, MAHSA University, Jenjarom 42610, Selangor, Malaysia
[f] Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Science Block, Alagappa University, Karaikudi-630 003, Tamil Nadu, India
[g] Department of Biotechnology, College of Engineering & Technology, SRM Institute of Science & Technology, Kattankulathur, Chennai, Tamil Nadu 603203, India
[h] Laboratory for Artificial Intelligence and Molecular Modelling, Center for Global Health Research, Saveetha Medical College, Saveetha Institute of Medical and Technical Sciences, Saveetha Nagar, Thandalam, Chennai, Tamil Nadu 602105, India

ARTICLE INFO

Keywords:
Drug Discovery
Drug Development
Artificial intelligence
Machine Learning
Deep Learning
Applications

ABSTRACT

Artificial intelligence (AI), particularly deep learning as a subcategory of AI, provides opportunities to accelerate and improve the process of discovering and developing new drugs. The use of AI in drug discovery is still in its early stages, but it has the potential to revolutionize the way new drugs are discovered and developed. As AI technology continues to evolve, it is likely that AI will play an even greater role in the future of drug discovery. AI is used to identify new drug targets, design new molecules, and predict the efficacy and safety of potential drugs. The inclusion of AI in drug discovery can screen millions of compounds in a matter of hours, identifying potential drug candidates that would have taken years to find using traditional methods. AI is highly utilized in the pharmaceutical industry by optimizing processes, reducing waste, and ensuring quality control. This review covers much-needed topics, including the different types of machine-learning techniques, their applications in drug discovery, and the challenges and limitations of using machine learning in this field. The state-of-the-art of AI-assisted pharmaceutical discovery is described, covering applications in structure and ligand-based virtual screening, *de novo* drug creation, prediction of physicochemical and pharmacokinetic properties, drug repurposing, and related topics. Finally, many obstacles and limits of present approaches are outlined, with an eye on potential future avenues for AI-assisted drug discovery and design.

1. Introduction

Early approaches for the drug discovery process might be 10–20 years of finding novel drug candidates or developing existing drug candidate processes. The main challenging issues of earlier drug discovery and development processes are time and cost. In this scenario, AI revolved around speeding up the possibility of drug discovery and development predictions with less cost [1]. Recently, AI has increased the productivity of the drug development process with the convergence of technologies such as biology, drug discovery, data analysis, machine learning, and deep learning. Knowledge of biology will help understand disease patterns to identify the target molecule, such as genes or proteins. AI's greatest advantages are processing the vast amount of stored target molecule data to analyse meaningful insights within a short time

* Corresponding author at: Laboratory for Artificial Intelligence and Molecular Modelling, Center for Global Health Research, Saveetha Medical College, Saveetha Institute of Medical and Technical Sciences, Saveetha Nagar, Thandalam, Chennai, Tamil Nadu 602105, India.
E-mail address: selnikraj@bioclues.org (C. Selvaraj).

https://doi.org/10.1016/j.ymeth.2023.09.010
Received 7 August 2023; Received in revised form 21 September 2023; Accepted 25 September 2023
Available online 29 September 2023

[2]. Machine learning and deep learning is a subfield of AI play a vital role in drug discovery and development. The machine learning is a process of previously stored statistical information of data and to provide future predictions and deep learning is a depth of knowledge discovered by artificial neural networks consists of several neurons with inputs layer, n-number of hidden layers, and an output layer, has the ability to solve the complex mathematical problems like human thinking [3]. AI models are becoming increasingly popular in drug discovery and development stages, especially for predicting biological or chemical properties. This is because AI models can be trained on large datasets of historical data and then use this data to make predictions about new compounds [4,5]. AI models' ability to make these predictions quickly and accurately is a significant advantage. Traditional drug discovery and development methods can be time-consuming, expensive, and often involve much trial and error. AI models can help to speed up the drug discovery process, and they can also help to reduce the risk of developing drugs that are ineffective or toxic [6,7].

1.1. AI in drug discovery pipeline

Artificial intelligence (AI) rapidly changes drug discovery and development by automating and speeding up tasks at every pipeline stage.

Target identification and validation: AI can discover and validate potential novel therapeutic targets by analyzing massive amounts of genomic and proteomic data. Artificial intelligence can be used to find proteins associated with a disease's pathway or altered genes [8].

Lead discovery: Artificial intelligence can be used to sift through huge chemical databases searching for promising new therapeutic leads. Artificial intelligence can also create novel pharmaceuticals with improved safety and efficacy [9].

Preclinical testing: Preclinical testing of new drug candidates can benefit from the application of AI to make predictions about their efficacy and safety. This can both speed up the medication development process and reduce the number of compounds that need to be tested in clinical trials [10].

Clinical trials: The success of clinical trials can be predicted, analyzed, and planned for with the help of AI. This can enhance the quality of clinical studies and speed up the process by which the most effective medications reach patients [11].

Regulatory approval: The chances of a medicine candidate being approved by regulators can be estimated with the help of AI. This can aid in deciding whether or not to move forward with clinical trials and speed up the drug development process [12].

Post-market surveillance: After a drug has been cleared for the market, AI can monitor it to ensure it doesn't cause any harm. This can aid in the early detection of any patient safety issues and the subsequent protection of the patients [13]. The U.S. Food and Drug Administration has acknowledged that AI-driven technologies have the potential to increase the reliability and efficacy of medical products but has stressed the need to safeguard patients during their creation and implementation. The Food and Drug Administration (FDA) intends to collaborate with industry to create regulations that promote innovation and protect patients.

The comparison between early approaches and AI-based drug discovery and development stages is described in Table 1.

Table 1
Comparison between the early and AI drug development process.

Process	Early drug development	AI drug development
Target selection & Target validation	The selection and validation identify the target molecules such as nucleic acid sequence (gene) or proteins.	AI analyses the open-source drug information bank. Which produces the score for the target drug to identify.
Compound Screening and lead optimization.	Over 5000–1000, compound screening identifies the hit to lead compound through chemistry, high-throughput screening, and virtual screening.	AI-based virtual screening is the compound database, which contains millions of compound information.
Preclinical studies –phase I	Laboratory tests for new compounds in in-vitro and in vivo to ensure efficacy and safety.	Machine learning approaches like supervised and unsupervised learning by in-vitro studies.
Clinical studies – phase II & III	This is a risky process for a new compound clinical trial with human participants.	The AI tool will help to identify the target molecule for potential therapy with the success rate of clinical trials.
Food and drug administration (FDA) approval		
Drug post-marketing		
Deliver drug molecules for target therapy		

1.2. Data representation

Data representation is a critical aspect of artificial intelligence (AI) in drug design, as it influences the accuracy and efficiency of predicting molecular properties, interactions, and drug candidates. There is a growing interest in developing new methods for representing data that can improve the performance of AI models [14]. Drug discovery using deep learning relies on chemical compounds and proteins as small molecule targets as the input data representation. Several representations for these molecules have been utilized in numerous machine learning models, which has substantially affected the accuracy of these forecasts [15]. Some of the data representations are listed in the Table 2.

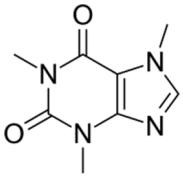
SMILES: In AI models for drug discovery and molecular property prediction, SMILES strings can be processed using techniques like tokenization, embedding, and recurrent neural networks (RNNs) to capture the molecular structure and predict properties like solubility, binding affinity, and toxicity [16]. SMILES-based models often leverage graph neural networks (GNNs) to capture molecules' graph-like nature and interactions. SMILES strings are easy to read and write, which makes them a convenient way to represent chemical data in AI models [17]. SMILES strings are extensible, which means that new features can be added to the format as needed. SMILES is a widely used format in the field of chemistry, which means that a large amount of data is available in this format.

Fingerprint: Fingerprint representation is a way of representing the chemical structure of a molecule as a series of numbers. These numbers can be used to represent the presence or absence of specific atoms or functional groups, or they can be used to represent the topological relationships between atoms in the molecule. Fingerprints are popular for representing chemical data in AI models because they are compact, efficient, and easy to interpret. They are also relatively insensitive to the order of atoms in a molecule, which makes them more robust to noise and errors in the data [18]. The choice of fingerprint type will depend on the specific application. For example, Morgan fingerprints are often used for virtual screening, while atom pairs fingerprints are often used for lead optimization. Fingerprint representations are used for various applications, including virtual screening (finding molecules with desired properties), quantitative structure–activity relationship (QSAR) modeling, and similarity searching in large chemical databases [19].

Molecular Graph: Molecular graph representation is a way of representing the chemical structure of a molecule as a graph. In this representation, the atoms in the molecule are represented as nodes in the graph, and the bonds between atoms are represented as edges [20]. Molecular graph representation is a powerful tool for AI drug design because it captures the topological relationships between atoms in a molecule. Molecular graphs represent atoms as nodes and bonds as edges in a graph, allowing machine learning models to leverage graph neural networks (GNNs) to process and predict various molecular properties [19].

Voxel: Voxel representation is a method of representing three-

Table 2
Molecular data representation for computational approaches.

Target Representation	Drug Representation	Description	Example
Sequence –Based Feature	SMILES	SMILES is Simplified Molecular Input Line Entry System it is most commonly used input to deep learning model. It represent the chemical structure used by computer can be easily learned sequenced based feature.	CC(=O) NC1 = CC = C (C = C1) O
	Fingerprint	It represent the molecule structure that convert into bit string i.e., presence of atoms indicate 1 and absence of atom indicate 0; it is useful method for describing the structural similarity of the molecule.	(1,1,0,0……0,1,0,1,1)
Structure-Based Feature	Molecular Graph	Molecular graph represents the chemical structure in terms of graph theory. It mapping of atoms constituting a molecule to node and chemical bonds to edges.	
	Voxel	Voxel is the combination of volume and pixel representation of three dimensions space. It is used in the target protein because it reacts with ligand instead of entire protein and it is very suitable for binding prediction.	3D target protein in cubes.

dimensional data as a grid of voxels. Each voxel in this representation is a 3D cube representing a small spatial volume. In the context of artificial intelligence models, voxel representations are commonly used for tasks that involve 3D data, such as medical imaging, computer graphics, and molecular modeling. Each voxel's value can indicate many data attributes such as color, intensity, or opacity. Voxel representation is a popular choice for expressing 3D data in AI models because it is a simple and efficient approach to representing the spatial relationships between items in the data. This data can be utilized to perform various tasks such as object detection, segmentation, and classification [19].

2. Artificial intelligence modeling in drug discovery

2.1. Structure-Based virtual screening methods with artificial intelligence approaches

Structure-based virtual screening (SBVS) is a computer method for identifying prospective medication candidates by screening a huge database of compounds against a known protein structure. SBVS is a strong technique that can be used to speed the drug development process. Artificial intelligence (AI) methods use machine learning and deep learning to improve the effectiveness of structure-based virtual screening. These approaches use machine learning to train scoring functions based on molecules' known binding affinities. AI-based scoring functions have been shown to be more accurate than traditional rule-based scoring functions, and they have been used to identify potential drug candidates for a number of diseases. SBVS has traditionally been regarded as a binary classification problem. Several studies have found that using dependencies between target classes in multilabel classification can enhance prediction accuracy [21]. For the purpose of drug design, the machine learning strategies can be classified into two main categories: supervised learning and unsupervised learning. In supervised learning, the machine learning model is trained using data that has already been annotated with the target result [22]. The model is trained to create a mapping between input and output. In unsupervised learning, a machine learning model is educated using data without labels. Without any input from the user, the model automatically discovers patterns in the data. The classifications are again sub-categories in to MLR: Multiple linear regression; PLS: Partial Least Squares; DT: Decision trees; RF: Random Forest, KNN: K-Nearest neighbours, MLP: Multilayer Perceptron; SVM: Support Vector Machines; SOM: Self-organizing Maps; PCA: Principal component analysis, as shown in the Fig. 1 [2,23].

2.2. Support Vector Machines

Support Vector Machines (SVMs) are powerful machine-learning algorithms commonly used in structure-based drug design. They are crucial in predicting molecular interactions, binding affinity, and other properties between ligands and target proteins. The training data preparation includes positive examples (ligands that bind to the target protein) and negative examples (ligands that do not bind), which are often labeled based on experimental binding data or known interactions [24]. SVM-based methods have been widely employed in developing

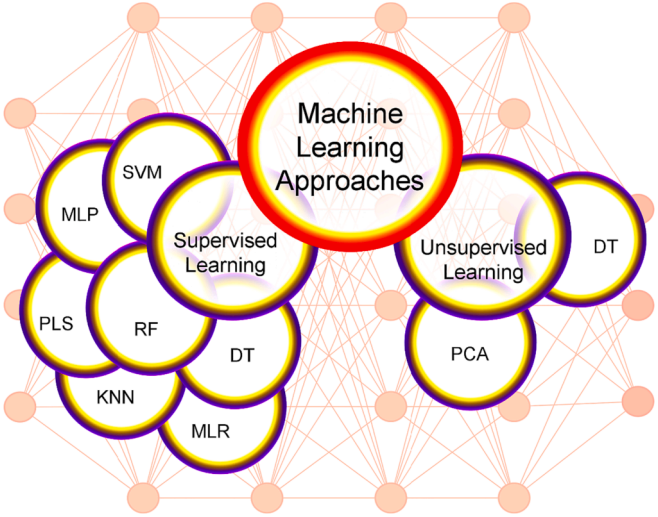


Fig. 1. Classification of Machine Learning strategies into their Supervised and Unsupervised learning methods.

target-specific scoring models known as SVM-SP (Support Vector Machine - Scoring Potential). SVM-SP models are used in molecular docking and virtual screening to predict the binding affinity or interaction energy between a ligand and a specific target protein (Fig. 2) [25]. These models enhance the accuracy of predicting ligand binding and assist in identifying potential drug candidates. The development of a general scoring function known as SVMGen (Support Vector Machine - General) that includes statistical pairwise potentials of docked protein–ligand pairs is a notable advancement in molecular docking and virtual screening. SVMGen leverages machine learning techniques, particularly SVMs, to predict protein–ligand complexes' binding affinity or interaction energy [26]. The incorporation of statistical pairwise potentials enhance the accuracy of binding affinity predictions. MIEC-SVM uses a combination of molecular interaction energy components (MIECs) and an SVM model to predict the binding affinity of molecules to target proteins [27]. MIECs are a way of representing the energetic interactions between atoms in a molecule. SVM methods can be quite helpful for various aspects of post-docking analysis in structure-based drug design. Post-docking analysis refers to examining and refining results obtained from molecular docking simulations, where the interactions between a ligand (small molecule) and a target protein are predicted [28].

2.3. Deep learning

Deep learning has profoundly influenced many disciplines, including structure-based approaches to drug design. Structure-based methods can use their three-dimensional structures to analyze and predict molecules and materials' properties, interactions, and behaviors [29,30]. The accuracy and efficiency of these strategies have been improved with the help of deep learning techniques [31]. Scalable three-dimensional protein structure prediction from amino acid sequences has been achieved by employing deep learning techniques. In recent years, deep learning approaches have shown much promise for solving this long-standing problem in computational biology [32,33]. Protein-ligand binding affinities have been predicted using deep learning techniques. Deep learning methods have been shown to be more accurate than conventional methods at this crucial stage of the drug discovery process. Methods like Generative adversarial networks (GANs) and variational autoencoders (VAEs) are helpful for structure-based drug discovery because they can generate novel molecular structures with the desired properties. In extension to docking methods, the deep learning methods accelerate molecular dynamics simulations by predicting potential energy landscapes using deep learning models, allowing the system to efficiently explore relevant configurations [34,35]. CNNs can be used to examine the trajectories generated by molecular dynamics simulations. They can clearly monitor the structural shifts, binding events, and other dynamic behaviors that would otherwise go undetected by more traditional techniques [36,37]. In drug discovery, deep learning methods are applicable to both academia and pharmaceutical industries, and this application is represented in Fig. 3. Deep learning architecture showing the input layer, and output layer in the external, while multiple hidden layers in the middle layer, and represented the input and output layer, that helps in the machine learning predictions as shown in the Fig. 4.

2.4. Bayesian networks

Bayesian networks (BNs) have several advantages over other machine-learning methods for drug discovery. First, BNs can represent data uncertainty. This is significant because the data used in drug discovery is frequently noisy and incomplete. Second, BNs are relatively simple to interpret. Bayesian networks, also called probabilistic graphical models, are statistical representations of probabilistic relationships between variables [38]. Structure-based drug designs embedded in Bayesian networks have been used to model and analyze complex interactions and dependencies within molecular systems. A molecular system's interactions between its parts, such as proteins and ligands, can be modeled using Bayesian networks. Atoms, residues, and ligands are represented by the graph's nodes, while their probabilistic dependencies and interactions are shown by the graph's edges [39]. It can predict a compound's activity based on its structural features and interactions with target molecules. They provide a more accurate estimate of a compound's activity by considering failure possibilities. Molecular descriptors and interaction data can be integrated into a single Bayesian network to help direct virtual screening. This is helpful for sorting and prioritizing compounds for further experimental testing. In iterative drug design, Bayesian networks can help by shedding light on how

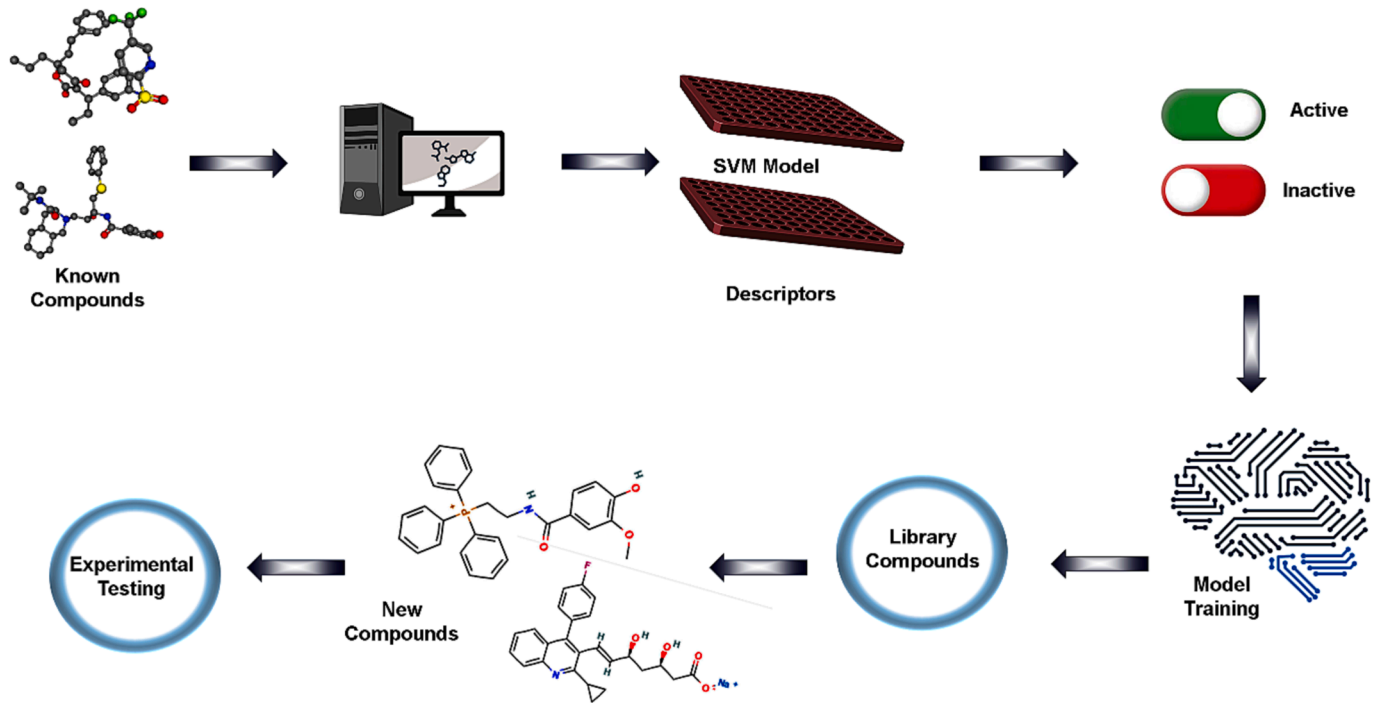


Fig. 2. Screening approach for the SVM-based algorithms in generating new compounds for the experimental testing.

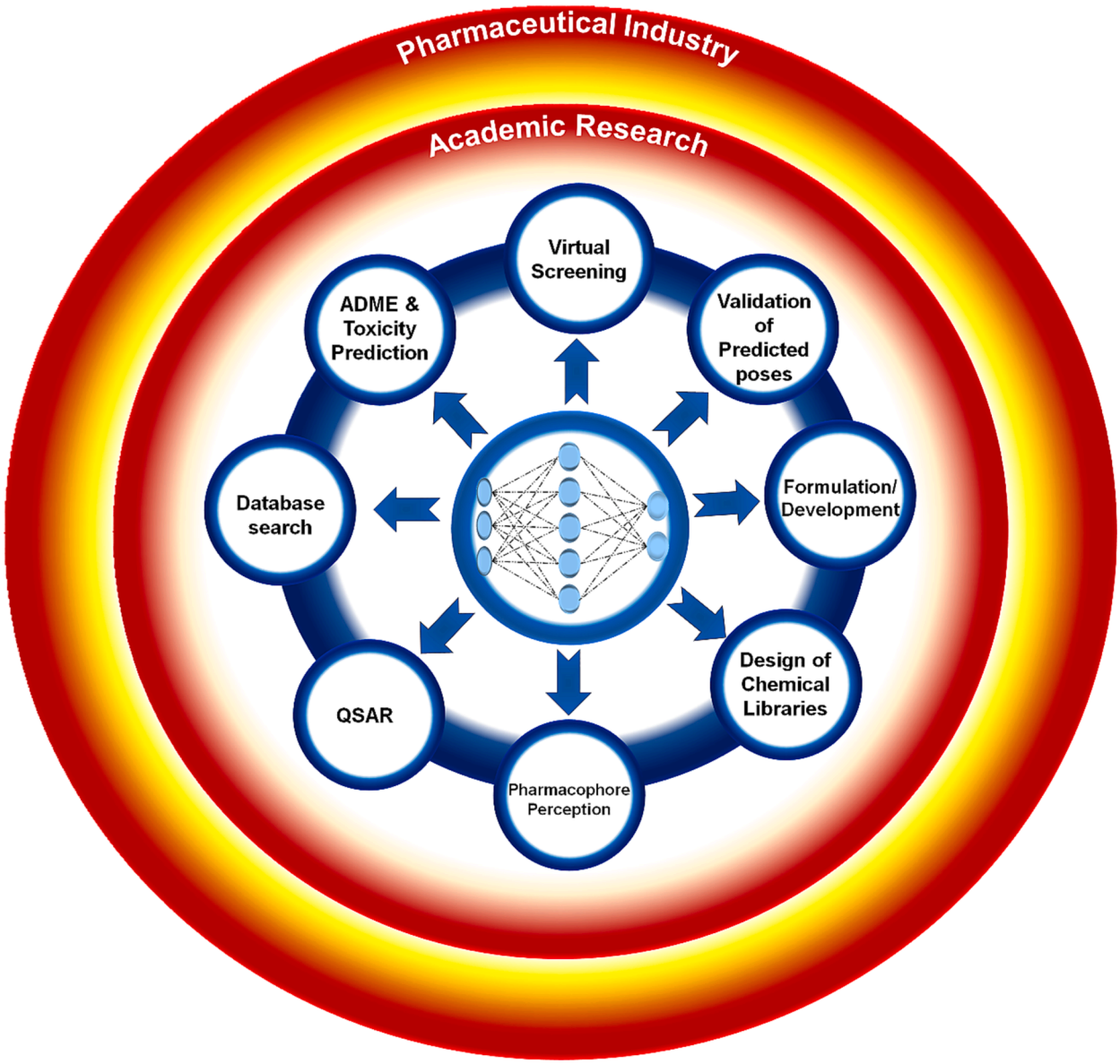


Fig. 3. Deep Learning method in molecular modeling applications, especially in increasing the efficiency of drug discovery in the academic and pharmaceutical industry.

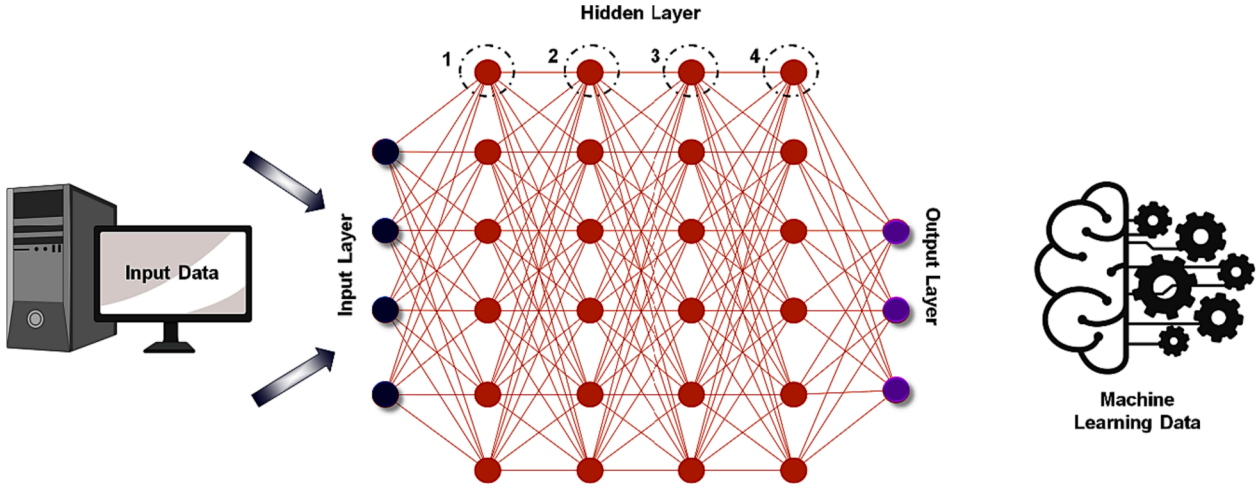


Fig. 4. A deep learning architecture's external input and output layers and the several hidden layers in the middle layer facilitate machine learning predictions.

changing individual molecular components or interactions might affect the final product. Overall, BNs are a promising new technology for structure-based drug discovery. They have several advantages over machine learning methods and are becoming increasingly popular in drug discovery. But, in addition to the Bayesian network, expert domain knowledge is crucial for ensuring that the network accurately captures the relevant relationships within the molecular system. In general, artificial intelligence techniques have had a major effect on structure-based drug design, which is the procedure of using molecular and structural information to create new drugs or enhance existing ones. Artificial intelligence methods can hasten the development of new medicines by making the various steps in the drug design pipeline more productive and precise [40]. Through these, the new potent molecules are yielded towards the active site as per the spatial requirements of the active site, as shown in Fig. 5, and these molecules are much more potent than the traditional screening method.

2.5. *Ligand-based virtual screening methods with artificial intelligence approaches*

In drug discovery, ligand-based virtual screening is a computational method to identify potential drug candidates from large compound libraries using a reference molecule's known properties. Finding compounds with structural and/or chemical features in common with a

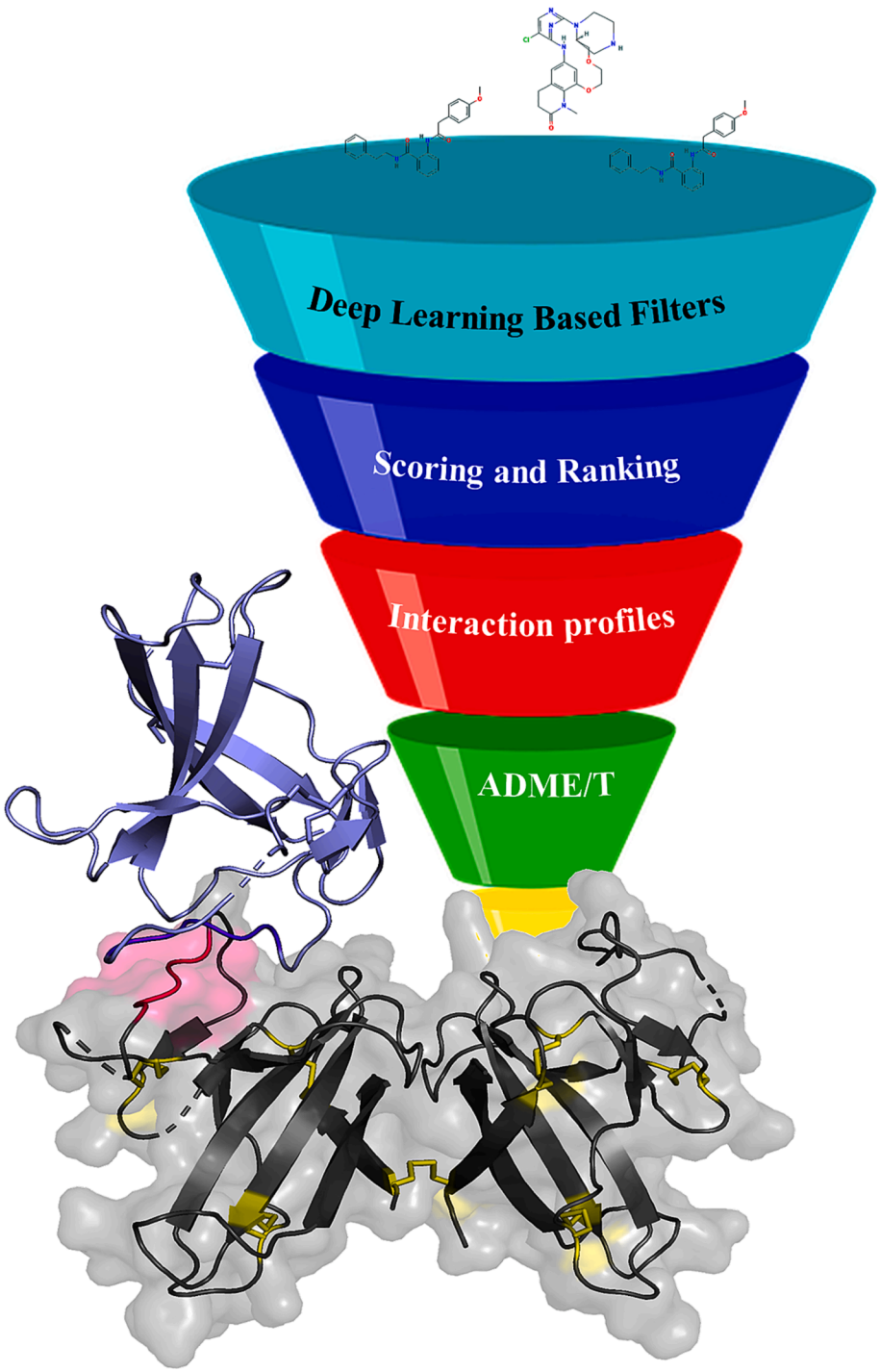


Fig. 5. Representation of deep learning methods in the structure-based methods, targeting the better-hit compounds with respect to the active site of the receptors.

known ligand is the goal of the LBVS method, which is helpful due to the lack of structure details of the receptor [41]. AI techniques are widely used in ligand-based drug design, which involves the optimization of small molecules to interact with target proteins or biomolecules. For correlating molecular descriptors (such as chemical properties) with biological activities, QSAR models are developed using AI techniques like machine learning and deep learning. These models support the understanding of structure–activity relationships and the prediction of new compound activity. AI techniques cluster and categorize ligands based on their structural and chemical similarities. This helps in identifying lead compounds and understanding chemical diversity. AI algorithms guide the selection of compounds for experimental testing by identifying the most informative ones for refining predictive models [42].

2.6. QSAR modeling – deep learning approach

QSAR models are useful in drug discovery because they can be used to pinpoint compounds that show promise [43]. QSAR is useful for predicting which compounds are more likely to be effective drugs by using a statistical model to correlate the molecular properties of a compound with its biological activity[44,45]. This can help researchers save time and money by narrowing their focus to the most promising compounds early in the drug discovery process. When experimental data and facilities are unavailable, QSAR is a highly effective method for discovering new compounds [46,47]. This is because chemical and structural features can be used in QSAR models to predict a compound's biological activity. This is a fast and low-expense technique to do compared to experimental methods. As a method for predicting how molecules will interact with biological targets, QSAR draws on organic chemistry [33,48]. It considers pharmacological information to evaluate the effectiveness and safety of a compound [46]. QSAR is the process of developing mathematical equations or models that relate molecular descriptors (features that describe the structure of a molecule) to biological activity. Based on their descriptors, these equations can be used to predict the activity of new compounds [49]. The values of biological activity of ligands serve as the dependent variable and the derived descriptors serve as the independent variables in multiple linear regression analyses of data gathered from the literature or own source of compounds. Using regression and classification on a large set of structure–property data, the empirical method known as QSAR attempts to demonstrate the link between biological activity and chemical structures. This model can be used to efficiently train and apply its predictions of unique chemical features and their biological capabilities without relying on time molecules. Because of this, QSAR has become increasingly popular and is being used in many different areas, such as drug design and toxicity prediction [35].

The applications of QSAR methods are enormous, but the main few are listed below.

- QSAR enables researchers to predict compounds' biological activity or properties without conducting time-consuming and extensive experimental tests. QSAR reduces the need to synthesize and test many molecules by focusing on more likely active compounds. This brings down the costs of chemical synthesis.
- QSAR rationally predicts Biological Activity, Physical Property, Chemical Property, Pharmaceutical Property, Toxicity, Environmental Behavior, Material Property, Food and Flavor, Agrochemical Property, Cosmetic and Personal Care Product and Property-Property Relationships activities/properties.
- QSAR uses big data to provide aggregate information and insights that aid in predicting various molecule activities and properties.
- The descriptors used in QSAR models can provide insights into the mechanism of action of compounds.

AI algorithms, particularly those based on machine learning and deep learning, have gained significant attention due to their efficiency in terms of time and cost. These algorithms can process and analyze large datasets much faster than traditional methods, and their predictive capabilities can potentially accelerate drug discovery and other research areas. Deep Neural Networks (DNNs) are a type of AI model that can predict molecular properties [50]. Concurrently training the generative and predictive stages of DNNs can introduce reinforcement learning, in which the output is rewarded or penalized based on a specific property. Training AI models at the same time can introduce bias into the output. This can be useful if the goal is to improve or degrade specific molecular properties [18]. QSAR and molecule design both benefit from the MMP algorithm. The method seeks to find pairs of molecules that differ in only one structural aspect but share another characteristic, such as charge or bioactivity. Assigning a single charge to a single lead molecule most likely refers to modifying a lead compound by adding or removing a functional group to change its bioactivities. Alterations can also be seen in other ML techniques, such as deep neural networks (DNNs), random forests (RFs), and gradient boosting machines (GBMs), in addition to MMP. It has been demonstrated that DNN is superior to RF and GBM in terms of prediction accuracy. Information on millions of compounds can be found in large databases such as PubChem and ChEMBL. This information includes details such as the compounds' structures and potential targets. The bioactivity of a drug, such as its intrinsic clearance, ADMET, oral exposure, and mechanism of action, can be predicted with MMP and ML. Toxicological optimization is an expensive and time-consuming step in drug discovery and this also performed using the AI pipeline as shown in the Fig. 6. However, this is an essential process that improves the final product. The capability of extracting additional information from an input database by means of mapping is not only the essential aspect of the feature but also the most advantageous aspect of the selection. Additionally, it has prepared network input variables by using the input database as a basis. In a later stage, when the input contains unrelated and redundant information, selecting the appropriate method to minimize the possibility of overfitting is essential [22,31].

2.7. Different models in QSAR

The QSAR approach that is standard is one that is simple and can be applied to substances that are chemically similar to one another. According to research that compared different molecular fields, the aesthetic appeal of the object increased the amount of molecular activity; as a result, the experiment was a huge success. The use of the atom or fragment model on example molecules has become increasingly common in recent years because of the model's growing complexity and the obscurity and precision of its descriptors. To learn the chemistry behind the performance of better-designed molecules, this method expands the range of prediction at the expense of interpretability [51]. QSAR modeling for a better pool requires the context-dependent selection of the most relevant subsets of descriptors. This improves models' generalizability and makes interpretation easier by eliminating unnecessary descriptors. There are generally two primary methods used to assess models. An initial step for ML is determining which features or descriptors are most important for enhancing molecule properties. Another method is appropriate for training the model on substructure-type descriptors to project the model's significant features and highlight the characteristics associated with the more favorable activity [52]. The molecular "heat map" comprises molecules with colored atoms based on their contribution. The modeled property is tested by applying the prediction of descriptors and model-independent approaches to feature interpretation to small changes to the input descriptors. Despite its flaws, this approach has become standard practice in the field of interpretation. Mechanistic interpretability is enhanced when the variables were chosen to have the same sign and size of their coefficient across the multiple QSARs. This highlights the importance of considering the statistical method that can distinguish correlation from causation and interpretation, which are not necessarily tied to a

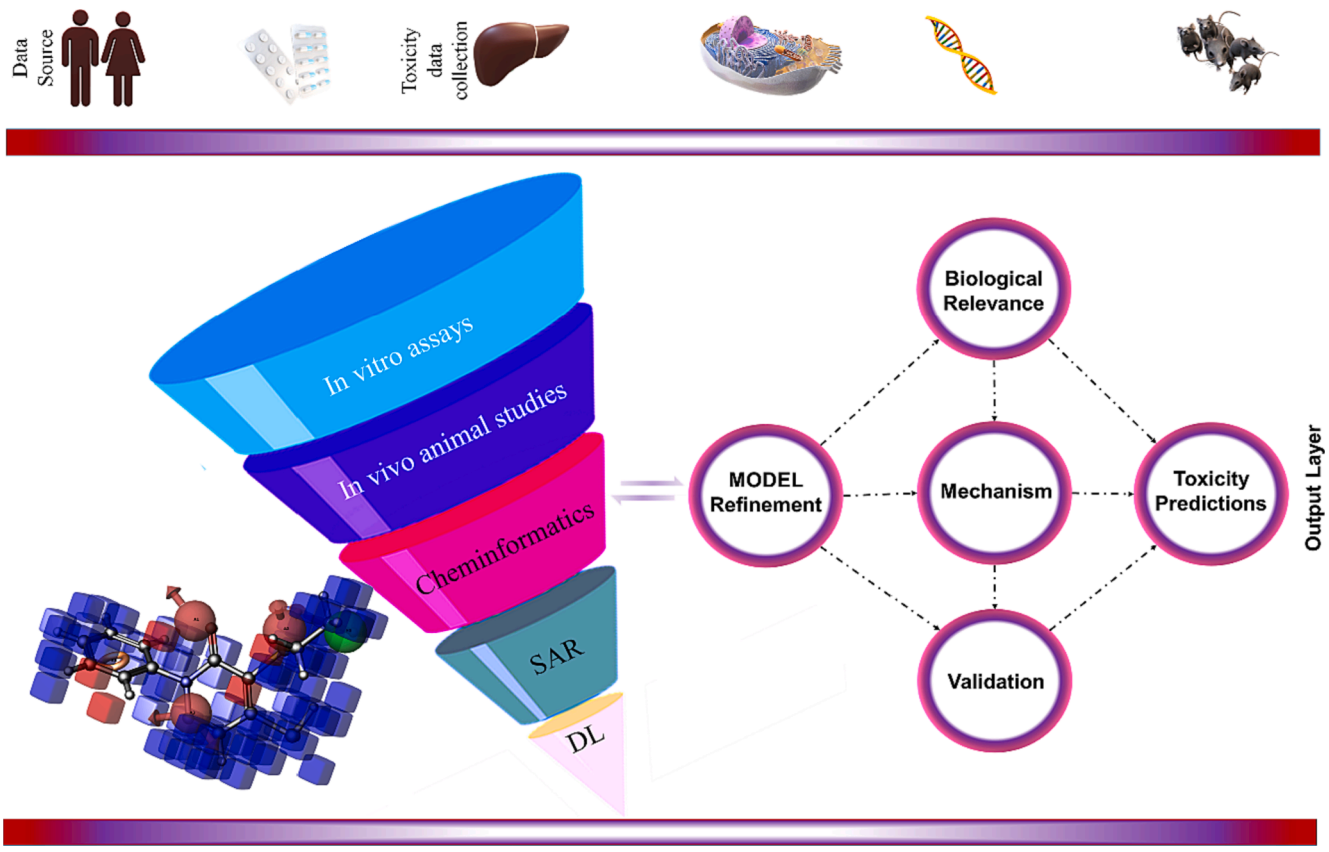


Fig. 6. Development of ML and DL-based toxicity predictions and ligand-based data collections for the outcome layer generation.

mechanism. When compared to the other prominent ML approaches, the DNN machine learning method is currently the most widely used ML method in QSAR. The application of DNN methods is extremely widespread and can be found in various fields [53]. The output of the DNN into QSAR as well as the numerous other ML methods utilized in QSAR modeling such as kNN, partial least squares (PLS), support vector machines, relevance vector machines, and relevance vector machines. DNN and ML are utilized in the pharmaceutical industry for accuracy prediction, sensitive tunable hyper-parameter, descriptor selection, enhanced training, and model interpretability. For many years, RF techniques were one of the most popular approaches in QSAR because they make meaningful predictions with adjustable parameters and can be parallelized. Furthermore, the degree of agreement in prediction between different RF trees can be used to calculate AD. Boosting is also an accurate and fast method, especially with the most recent implementation and light gradient boosting machine [54].

2.8. ANN-QSAR

The use of ANN, a popular ML algorithm, has increased for the purpose of developing QSAR models. Artificial neural networks (ANNs) are a subset of machine learning (ML) that were created as a direct analog to biological NNs. Three main components go into making up a typical ANN: the transfer function, the learning rule, and the connection formula. It has been found that feed-forward artificial neural networks (FF-ANNs) are the most used type of ANNs in the real world [53]. The most popular method for estimating FF-ANN parameters is the BP approach, which uses derivatives of the error function to minimize the network error and, thus find the optimal parameters. The BP algorithm's goal is to minimize the discrepancy between the expected and observed values, whether those values are the output or anticipation of a process [55].

2.9. ML models – QSAR

In recent years, molecular design and screening, chemical structure prediction, and category forecasting are just a few areas where ML has been increasingly applied in QSAR models. ML family methods like SVM and RF are frequently used when looking for a new and effective treatment. Produced via graph evaluation of a collection of 2D or 3D chemical descriptors, chemical fingerprints find application in a wide range of ML models and prediction tasks. Gene sequencing, single-cell sequencing, multi-omics interaction data, protein structure, and gene-protein interactions have all taken significant leaps forward thanks to the incorporation of massive datasets and ML [56]. To make this prediction, ML models are used, and these models are further subdivided into discriminating prediction and generative prediction variants. This is because a common application of ML models is to predict the efficacy of candidate compounds, whereas traditional QSAR primarily focuses on this ability. Using discriminative models, which map the association between the sample's characteristics and the prediction's target based on the data, is related to pattern recognition. These models are used in the process. The discriminative model is used to evaluate the distribution of the sample data, whereas the generative model is used to forecast new compounds based on the learned distribution. Both models are used in conjunction with one another. Both models are utilized together to achieve the desired results. Some common ways of categorizing these methods are supervised, semi-supervised, unsupervised, or self-supervised learning. This model can only be used for supervised learning if discriminant models are employed, as its unique features make it inappropriate for unsupervised learning. Feature selection, algorithm quality, and training set quality are typical applications of these models [57]. Conformation-related indicators in QSAR and conformational calculation for isomers were used to evaluate the research using quantum chemistry, semi-empirical approaches, and molecular dynamic

simulation. The ML-based QSAR method is used to predict how the solvent environment affects conformational changes and molecular energy by evaluating the energy changes that occur during conformational changes. Since ML is data-driven, it only needs to use standard QSAR algorithms to learn from the training set after all relevant features pertaining to the nature of the QSAR have been exhausted. The fundamental characteristics shift based on the structure–activity or prediction objective. The study of physical and chemical properties that are difficult to compute using theoretical methods like density functional theory or MD but can typically be computed by cheminformatics models is greatly facilitated by ML-based methods [16].

2.10. DL – QSAR

Success with DL in imaging techniques has led to its application in other areas of medicine, such as machine translation and speech analysis, and now DL is also being used to study genetic variation in populations, create synthetic biology drugs, and improve disease diagnosis. Different types of DNNs and CNNs, as well as multitask learning, capsule networks, self-encoding decoders, GANs, long short teams, variational AutoEncoders, and so on, are all examples of DL methods. Data in the biological sciences are extremely rich in information, which can be represented structurally through multitask learning. For instance, a GAN can model and generate microbial metabolic networks and loop networks. Combining Variational AutoEncoder and GAN for learning makes it possible to encode complex structures. The protein structure can be predicted by DL from the gene sequence alone. A convolutional neural network (CNN) is used to do learning based on the force field and the protein–ligand complex, while atoms, strategy networks, and Monte Carlo trees are used to predict the chemical reaction. The stated issue is addressed better by the DL approach in QSAR, which includes CNNs, DNNs, and NNs with more than two layers and many neurons, as well as the deep architecture of the deep belief network (DBN) used to fine-tune the network's initial parameter and reduce its associated bias. Higher-level features in DBN are generated from lower-level ones, making it a form of generative unsupervised learning. Similar to how a deep belief network (DBN) applies an unsupervised learning algorithm layer by layer, an autoencoder's two main parts are its encoding of the input data and its decoding of the hidden units to reconstruct the input data, so the number of input and output unsupervised learning layers is equal to the number of input and output hidden units. When it comes to pretraining the NN, the autoencoder training algorithms are used as a crucial component. As a result, it greatly reduces redundancy issues like getting stuck in a local minimum and speeds up the model generally. The NN method is described by CNN in terms of weighted and biased neurons. For example, even locally stored databases with a small number of descriptors can tap into the network's full potential. However, the goal of the connected redundancy problem is quickly achieved in the case of a large data set with a significant number of descriptors, like the ChEMBL database, and network parameters that are quickly assayed and rapidly lead to overfitting. Unlike a feed-forward, a NN employs all of the connected layers. CNN's architecture consists of three layers: the conventional layer, the pooling layer, and the fully connected layer. The topmost layer is the fully linked layer. After that, each map is sub- and down sampled using average sum or max pooling over pxp neighbor pixels, where p denotes the difference between the small and large input data. The conventional and connected layers are used to extract features from neurons, and it has K-filters and convNet with small-size data [58].

2.11. Decision tree algorithms

Decision trees are an ML algorithm that, when working with tabular data, is competitive with NNs in terms of quality and performance and, in many cases, outperforms NNs. A decision tree is a useful tool for decision analysis because it provides a clear and concise representation of decisions and the decision-making process. A decision tree is a form of supervised learning. Like other supervised learning algorithms, decision trees can be used to solve regression and classification problems. The target chemical's class or value can be predicted using the decision tree's training model by applying a set of rules derived from the data used to train the model. Many aspects of ML, such as classification and regression, have been influenced by trees because of their numerous real-world analogs. The decision tree begins its analysis at the node representing the tree's root to make an educated guess on the record's class label [59]. Check whether the value of the root attribute is the same as the value of the attribute in the record. Trees have influenced many areas of machine learning due to their abundance of real-world analogs, particularly in classification and regression. The root node of the decision tree is where the process of determining the record's class label begins. Verify that the attribute value in the record corresponds to the root attribute value. The discrete or continuous nature of the data being worked with determines the problem-solving approach taken by the method. Trees have influenced many areas of machine learning due to their abundance of real-world analogs, particularly in classification and regression. The root node of the decision tree is where the process of determining the record's class label begins. Verify that the attribute value in the record corresponds to the root attribute value. If the value you compared it to differs from what you expected, take the branch to the next node. The discrete or continuous nature of the data being worked with determines the problem-solving approach taken by the method on the nature of the data being worked with, it can solve problems in either the discrete or continuous setting. Due to the categorical nature of the data, the ID3 method can only be replicated using the WEKA tool. This is significant in terms of the features of decision trees. ID3′s simulation environment does not support continuous data sets. Both CART and C4.5 share some of ID3′s characteristics. C4.5 is superior to CART because it allows continuous data sets to be used in simulations, whereas CART does not. The decision tree lays out all the options clearly and displays their progression in a single diagram, making it easy to compare and contrast the various possibilities. One of its benefits is that it is open and honest. Selecting more biased physiologies and their intuitive nature are two additional benefits. The ease of categorization and interpretation is another plus. Decision trees can achieve excellent results by using variable screening and feature sections. There is zero impact of non-linearity on the performance characteristics of the decision tree. Decision tree techniques are used to classify the characteristics into different buckets to establish whether a split is the "best" option. Because each branch must use the same criterion for splitting, the resulting partition is as clean as practically possible [60].

2.12. Random Forest

Random Forest (RF) is an ensemble learning method that uses multiple decision trees to produce a more reliable and accurate model. Particularly useful for classification and regression. Since RF has proven effective and useful in many contexts, it has quickly become the most prominent ML algorithm. RF is more accurate, and with this, it can offer much higher accuracy on a wide range of data sets, and it is robust to noise and outliers, as it is not sensitive like other machine learning algorithms. In addition, it is easy to understand and interpret for beginners, and its versatile nature provides applicability for various tasks, including classification, regression, clustering, and feature selection [61]. Leo Breiman proposed the Random Forest (RF) algorithm in 2001. Breiman's original paper, "Random Forests," introduced the concept of ensemble learning using decision trees. Due to limitations in categorizing and computing important metrics, RF is used for quantile prediction, survival analysis, and causal inference. It is also used in agriculture, land cover classification, remote sensing, ecology, wetland classification, bioinformatics, genomics, and QSAR. The success of Random Forest and its derivatives in these applications can be attributed to its ability to handle non-linear relationships, capture interactions between multiple

variables, and handle noisy or incomplete data [62]. It's important to note that while Random Forest is powerful, its performance depends on proper data preprocessing, feature engineering, and hyperparameter tuning to suit the specific water resource application.

2.13. Software available for QSAR

A limited collection of well-known databases, websites, and software tools, both free and commercial, that can be utilized in QSAR research is provided in Table 3.

2.14. Application of AI in pharmaceutical industries

Applications of AI in drug discovery, development, manufacturing, and marketing are reshaping the pharmaceutical industry at a rapid pace. Especially, the pharma industry drug discovery process is being sped up with the help of AI by means of the detection of novel drug targets, the creation of novel molecules, and the prediction of the safety and efficacy of these drugs. It is highly likely that the application of this technology will have an even more significant impact on the pharmaceutical industry as the field of artificial intelligence continues to progress. Finding new drugs, developing new drugs, and manufacturing drugs are all processes that artificial intelligence could completely revolutionize. This includes the finding of new drugs and the development of new drugs. This could lead to the development of new treatments for diseases as well as improved patient outcomes. Applying artificial intelligence could fundamentally change these procedures and enhance decision-making throughout the drug lifecycle. Some of the AI tools are available online for making the drug discovery in more accurate and ease are provided in the Table 4. It's important to note that AI tools in drug discovery often require access to large and diverse datasets, as well as validation through experimental testing.

2.15. Blockchain technology integrated AI for drug discovery

The pharmaceutical sector and the process of creating new medicines could benefit greatly from the use of blockchain technology. The decentralized and immutable ledger architecture of blockchain can improve the security and integrity of sensitive drug-related data [63]. This is critical in medication design, because data accuracy and privacy are critical. Researchers can safely store and share data without fear of illegal access or data modification[64]. Clinical trial data management can be streamlined using blockchain. A blockchain can record patient permission, data collection, and results, assuring transparency and preventing data tampering[65,66]. This has the potential to improve the trustworthiness of clinical trial results. While the potential benefits of blockchain technology in medication design and the pharmaceutical business are substantial, it is crucial to stress that adopting blockchain solutions in a highly regulated and complex subjects like healthcare and drug research is fraught with difficulties. Scalability, interoperability with existing systems, regulatory compliance, and industry-wide adoption requirements are all hurdles[67,68]. Nonetheless, as blockchain technology matures, its uses in drug creation are anticipated to expand, opening up new avenues for creativity and data security in the area. By combining the transparency, security, and collaboration capabilities of blockchain with the data analysis and prediction capabilities of AI, the drug discovery process can become more efficient, cost-effective, and trustworthy[69,70].

2.16. Limitations of AI in pharmaceutical industries

Meanwhile, the AI and AI-based technology revolution has created constraints in the pharmaceutical industry. In the pharmaceutical industry, where data is often siloed and difficult to access, the uncertainty of the quantity and quality of data poses a significant challenge when training AI models on these smaller datasets. In addition, AI models are

Table 3
QSAR programs available for Ligand-based approaches.

S. No	Program	Named for	Summary and Link
1.	ECOSAR	Ecological Structure Activity Relationships	The United States Environmental Protection Agency (EPA) and Syracuse Research Corporation created this software to foretell the aquatic toxicity of industrial chemicals. https://www.toxit.it/en/services/software/ecosar
2.	ChemACE	The Chemical Assessment Clustering Engine	It's made to help with read across and data gap filling for untested substances, as well as reviewing and prioritizing large chemical inventories. https://www.epa.gov/tsca-screening-tools/chemical-assessment-clustering-engine-chemace
3.	ChemSTEER	Chemical Screening Tool for Exposures and Environmental Releases	It can calculate hazards to workers and the environment from chemical production and use. https://www.epa.gov/tsca-screening-tools/chemsteer-chemical-screening-tool-exposures-and-environmental-releases
4.	EPI Suite™	Estimation Programs Interface	Predicts the physical/chemical property and environmental fate estimation programs https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface
5.	QSPR-Thesaurus	CAse Studies on the Development and Application of *in Silico* Techniques for Environmental Hazard and Risk Assessment	It exemplify REACH-related hazard assessments for four classes of chemical compound, namely, polybrominated diphenylethers, per and polyfluorinated compounds, (benzo)triazoles, and musks and fragrances. https://cadaster.eu/node/118.html
6.	ReachScan	Estimate the Surface water chemical concentrations in drinking water utilities downstream from industrial facilities	ReachScan will estimate chemical concentrations in single or multiple stream (segments) reaches by simple dilution or using simple fate algorithms. https://www.epa.gov/tsca-screening-tools/reachscan-exposure-assessment-model
7.	OSIRIS property Explorer	Drawing the Chemical Structure	The OSIRIS Property Explorer lets you draw chemical structures and calculates on-the-fly various drug-relevant properties whenever a structure is valid. https://www.organic-chemistry.org/prog/peo/
8.	OECD QSAR Toolbox	QSAR Tool Box	Toxicological predictions based on qualitative and quantitative structure–activity relationship methods, including read-across, are

(*continued on next page*)

Table 3 (*continued*)

S. No	Program	Named for	Summary and Link
			made accessible to the user in a clear and understandable format by the OECD QSAR Toolbox. https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm
9.	ToxPredict	Toxicity Prediction	Several (Q)SAR models were built with OPENTOX for several REACH endpoints (carcinogenicity, mutagenicity, aquatic / fish toxicity, LogP). https://old.opentox.org/dev/testing/testcasedevelopment/ toxpredict
10.	VEGA	Virtual models for property Evaluation of chemicals within a Global Architecture	Property evaluation for chemicals based on the assessment values. https://www.vegahub.eu/portfolio-item/vega-qsar/
11.	ToxTree	Toxicity Data	Toxtree is a feature-rich, adaptable, and user-friendly open-source program that uses a decision tree methodology to estimate toxic hazard. https://toxtree.sourceforge.net/
12.	QSAR TOOLBOX	QSAR Tool Box	The purpose of QSAR Toolbox is to help governments, the chemical industry, and other interested parties fill in the gaps in (eco)toxicity data necessary for evaluating the risks posed by chemicals. https://qsartoolbox.org/
13.	Dragon	Descriptor based QSAR	DRAGON provides more than 1,600 molecular descriptors divided into 20 logical blocks to help the user to manage multiple descriptors. https://www.talete.mi.it/products/dragon_description.htm
14.	Phase	ATOM and Pharmacophore based 3D QSAR	The PHASE pharmacophore and 3D QSAR models allow for the efficient extraction of actives from a prototypical database, the rationalization of structure–activity data, and the prediction of the activity of new compounds. https://www.schrodinger.com/products/phase
15.	VolSurf+	ADME and Pharmacokinetic modeling	From the 3D Molecular Interaction Fields (MIFs) generated by the GRID software, VolSurf + generates 128 molecular descriptors that are both relevant to ADME prediction and easy to interpret. https://www.moldiscovery.com/software/vsplus/

predictive and have the potential to be biased, which can lead to inaccuracies that, despite their apparent insignificance, can have major repercussions. Concerns are being raised by regulatory bodies in the pharmaceutical industry about the use of artificial intelligence in

Table 4
Name of the AI tools that supports modern Drug Discovery.

S. No	Name of AI Tool	Application	Website
1	AtonNet	Providing advanced computing facility with the incorporation of AI	https://atos.net/en/artificial-intelligence
2	DeepTox	Pipeline for predicting toxic effects of chemical compounds	https://bioinf.jku.at/research/DeepTox/tox21.html
3	DeepChem	Python library for machine learning and deep learning on molecular and quantum datasets	https://deepchem.io/
4	DeepPurpose	Deep Learning Based Drug Repurposing and Virtual Screening Toolkit (using PyTorch).	https://deeppurpose.readthedocs.io/en/latest/
5	Chemical VAE	AI tool for machine learning of molecular properties	https://github.com/aspuru-guzik-group/chemical_vae
6	DEEPConV-DTI	Neural Networks for Drug-Target Interaction prediction	https://github.com/GIST-CSBL/DeepConv-DTI
7	Deep Screening	For constructing deep learning models using public dataset or user provided dataset to search new compounds	https://deepscreening.xielab.net/
8	ODDT	Machine learning scoring functions (RF-Score and NNScore) to develop CADD pipelines	https://github.com/oddt/oddt
9	AIDDISON™	An integrated and easy-to-use tool for lead identification that brings together a suite of tools for modeling, docking and scoring molecules	https://www.sigmaaldrich.com/IN/en/
10	AMPL	Software pipeline for building and sharing models to further in silico drug discovery	https://github.com/ATOMScience-org/AMPL#AMPL-Features
11	AlphaFold	The solution for the protein folding problem	https://www.deepmind.com/research/highlighted-research/alphafold

adhering to regulatory guidelines and approval standards. This contributes to an increase in the level of complexity associated with the approval process for new drugs. Interpreting certain types of artificial intelligence models, particularly deep learning models, can be difficult. As a result of this lack of interpretability, it can be difficult to understand how AI arrived at a particular decision or recommendation, which makes it difficult for regulators, healthcare professionals, and patients to trust AI-based solutions and adopt them. Concerns about patient privacy, informed consent, and the possibility of discrimination are raised in relation to artificial intelligence (AI). The gathering of patient information for the purpose of gaining AI-driven insights must be carried out in a responsible and transparent way. In many cases, AI models possess different in-depth domain expertise than pharmaceutical researchers and clinicians. This can result in solutions being generated by AI that are correct from a technical standpoint but have no clinical relevance or applicability in the scientific world. Moreover, investing in computational resources and knowledgeable personnel to create and maintain AI systems can be a financial constraint. Due to limited resources, smaller pharmaceutical companies may struggle to adopt AI to the trends

followed by giant large cap pharma companies, which may create technology gaps.

2.17. Future potential of AI for the pharmaceutical sectors

Artificial intelligence (AI) has tremendous transformative potential in the pharmaceutical industry and is poised to revolutionize many facets of the industry, including drug discovery, development, manufacturing, and patient care. More advanced forms of AI will continue to revolutionize the pharmaceutical industry by facilitating more rapid, efficient, and patient-centered drug development and healthcare delivery. In the future, the addition to accelerating the identification of potential therapies, AI can fasten the drug discovery process by accurately predicting drug candidates and optimizing their properties. Personalized treatment approaches based on individual genomics and health records will be made possible by AI's ability to analyze such large amounts of genomics data. It has the potential to improve patient recruitment, trial design, and monitoring, all of which contribute to faster, less expensive clinical trials. In addition, AI models can improve their accuracy over time by continually learning from new data and adjusting to advances in medical and scientific understanding. However, maximizing the benefits of AI in the pharmaceutical industry will require resolving issues of data privacy, bias, regulatory compliance, and transparency. In addition to these factors, scientific expertise needs to advance and update to compete with AI trained models, and a failure to integrate AI methods in the pharmaceutical industry could result in the loss of jobs, and this may be highly witnessed in the future.

3. Conclusion

The rapid development of AI technology has been catalyzed by advancements in AI algorithms, particularly in deep-learning approaches. These advancements, along with increasing architectural hardware specialization (such as GPUs, TPUs, and large-scale parallel computing), and the availability of big data, have also contributed to the acceleration of this development. The success of recent efforts in natural language processing, image and voice recognition, and other areas has brought the topic to the attention of a wider audience, which has led to an increase in optimism. In many domains, artificial intelligence has already surpassed the performance of human specialists. Despite the progress made so far, we should approach the use of AI in drug discovery with an open mind due to the many obstacles that still need to be overcome. In particular, the acquisition of sufficient, high-quality, problem-specific data remains a significant barrier to the success achieved in other fields where AI has been applied and a major challenge in AI-assisted drug discovery. It will take the combined efforts of researchers working on artificial intelligence, pharmaceutical industry professionals, government regulators, and policymakers to overcome these challenges. Integrating artificial intelligence (AI) in a responsible and fruitful manner into the pharmaceutical industry requires forethought, rigorous validation, and an ethical framework. Artificial intelligence is still a relatively new technology; however, it has the potential to revolutionize the process of drug discovery and development. In the years to come, we can anticipate seeing an even greater variety of cutting-edge applications of AI technology in this sector as the underlying technology for AI continues to advance.

CRediT authorship contribution statement

CS and PNS: Concept, Original draft, and Supervision. CS and SKS: Figure drawing and refinement. BS and SKS: Data Collection. NKBD, BV, ASA, ESA, SZS and DV: Review and Editing.

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

Acknowledgements

CS thankfully acknowledges the Saveetha Medical College, SIMATS for providing the infrastructure facility to conduct this research work. The author would like to thank Deanship of Scientific Research at Majmaah University for supporting this work.

References

[1] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade, Artificial intelligence in drug discovery and development, Drug Discov. Today 26 (1) (2021) 80–93.
[2] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (6) (2019) 463–477.
[3] S. Dara, S. Dhamercherla, S.S. Jadav, C.M. Babu, M.J. Ahsan, Machine Learning in Drug Discovery: A Review, Artif. Intell. Rev. 55 (3) (2022) 1947–1999.
[4] L.K. Vora, A.D. Gholap, K. Jetha, R.R.S. Thakur, H.K. Solanki, V.P. Chavda, Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design, Pharmaceutics 15(7) (2023).
[5] H. Bao, J. Zhao, X. Zhao, C. Zhao, X. Lu, G. Xu, Prediction of plant secondary metabolic pathways using deep transfer learning, BMC Bioinf. 24 (1) (2023) 348.
[6] K.K. Mak, M.R. Pichika, Artificial intelligence in drug development: present status and future prospects, Drug Discov. Today 24 (3) (2019) 773–780.
[7] K. Wang, M. Li, Fusion-Based Deep Learning Architecture for Detecting Drug-Target Binding Affinity Using Target and Drug Sequence and Structure, IEEE J Biomed Health Inform PP (2023).
[8] F.W. Pun, B.H.M. Liu, X. Long, H.W. Leung, G.H.D. Leung, Q.T. Mewborne, J. Gao, A. Shneyderman, I.V. Ozerov, J. Wang, F. Ren, A. Aliper, E. Bischof, E. Izumchenko, X. Guan, K. Zhang, B. Lu, J.D. Rothstein, M.E. Cudkowicz, A. Zhavoronkov, Identification of Therapeutic Targets for Amyotrophic Lateral Sclerosis Using PandaOmics - An AI-Enabled Biological Target Discovery Platform, Front. Aging Neurosci. 14 (2022), 914017.
[9] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R.K. Ambasta, P. Kumar, Artificial intelligence to deep learning: machine intelligence approach for drug discovery, Mol. Divers. 25 (3) (2021) 1315–1360.
[10] Y. You, X. Lai, Y. Pan, H. Zheng, J. Vera, S. Liu, S. Deng, L. Zhang, Artificial intelligence in cancer target identification and drug discovery, Signal Transduct. Target. Ther. 7 (1) (2022) 156.
[11] G. Li, P. Lin, K. Wang, C.C. Gu, S. Kusari, Artificial intelligence-guided discovery of anticancer lead compounds from plants and associated microorganisms, Trends Cancer 8 (1) (2022) 65–80.
[12] A. Nayarisseri, R. Khandelwal, P. Tanwar, M. Madhavi, D. Sharma, G. Thakur, A. Speck-Planche, S.K. Singh, Artificial Intelligence, Big Data and Machine Learning Approaches in Precision Medicine & Drug Discovery, Curr. Drug Targets 22 (6) (2021) 631–655.
[13] N. Nagarajan, E.K.Y. Yapp, N.Q.K. Le, B. Kamaraj, A.M. Al-Subaie, H.Y. Yeh, Application of Computational Biology and Artificial Intelligence Technologies in Cancer Precision Drug Discovery, Biomed Res. Int. 2019 (2019) 8427042.
[14] S. Vatansever, A. Schlessinger, D. Wacker, H.U. Kaniskan, J. Jin, M.M. Zhou, B. Zhang, Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions, Med. Res. Rev. 41 (3) (2021) 1427–1473.
[15] C. Cerchia, A. Lavecchia, New avenues in artificial-intelligence-assisted drug discovery, Drug Discov. Today 28 (4) (2023), 103516.
[16] V.D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A.G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco, G. Melagraki, Advances in de Novo Drug Design: From Conventional to Machine Learning Methods, Int. J. Mol. Sci. 22 (4) (2021) 1676.
[17] X. Li, Y. Xu, H. Yao, K. Lin, Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors, J Cheminform 12 (1) (2020) 42.
[18] A. Ilnicka, G. Schneider, Compression of molecular fingerprints with autoencoder networks, Mol Inform 42 (6) (2023) e2300059.
[19] L. David, A. Thakkar, R. Mercado, O. Engkvist, Molecular representations in AI-driven drug discovery: a review and practical guide, J Cheminform 12 (1) (2020) 56.
[20] K. Amara, R. Rodriguez-Perez, J. Jimenez-Luna, Explaining compound activity predictions with a substructure-aware loss for graph neural networks, J Cheminform 15 (1) (2023) 67.
[21] S.M. Ruatta, D.N. Prada Gori, M. Flo Diaz, F. Lorenzelli, K. Perelmuter, L. N. Alberca, C.L. Bellera, A. Medeiros, G.V. Lopez, M. Ingold, W. Porcal, E. Dibello, I. Ihnatenko, C. Kunick, M. Incerti, M. Luzardo, M. Colobbio, J.C. Ramos, E. Manta, L. Minini, M.L. Lavaggi, P. Hernandez, J. Sarlauskas, C.S. Huerta Garcia, R. Castillo, A. Hernandez-Campos, G. Ribaudo, G. Zagotto, R. Carlucci, N.

S. Medran, G.R. Labadie, M. Martinez-Amezaga, C.M.L. Delpiccolo, E.G. Mata, L. Scarone, L. Posada, G. Serra, T. Calogeropoulou, K. Prousis, A. Detsi, M. Cabrera, G. Alvarez, A. Aicardo, V. Araujo, C. Chavarria, L.P. Masic, M.E. Gantner, M. A. Llanos, S. Rodriguez, L. Gavernet, S. Park, J. Heo, H. Lee, K.H. Paul Park, M. Bollati-Fogolin, O. Pritsch, D. Shum, A. Talevi, M.A. Comini, Garbage in, garbage out: how reliable training data improved a virtual screening approach against SARS-CoV-2 MPro, Front. Pharmacol. 14 (2023) 1193282.

[22] J. Scantlebury, L. Vost, A. Carbery, T.E. Hadfield, O.M. Turnbull, N. Brown, V. Chenthamarakshan, P. Das, H. Grosjean, F. von Delft, C.M. Deane, A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening, J. Chem. Inf. Model. 63 (10) (2023) 2960–2974.

[23] C. Selvaraj, I. Chandra, S.K. Singh, Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries, Mol. Divers. 26 (3) (2022) 1893–1913.

[24] S. Limbu, S. Dakshanamurthy, A New Hybrid Neural Network Deep Learning Method for Protein-Ligand Binding Affinity Prediction and De Novo Drug Design, Int. J. Mol. Sci. 23 (22) (2022) 13912.

[25] K. Sinha, S. Parwez, S. Mv, A. Yadav, M.I. Siddiqi, D. Banerjee, Machine learning and biological evaluation-based identification of a potential MMP-9 inhibitor, effective against ovarian cancer cells SKOV3, J. Biomol. Struct. Dyn. (2023) 1–19.

[26] G. Li, J. Li, Y. Tian, Y. Zhao, X. Pang, A. Yan, Machine learning-based classification models for non-covalent Bruton's tyrosine kinase inhibitors: predictive ability and interpretability, Mol. Divers. (2023).

[27] S. Zhong, X. Guan, Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties, Environ. Sci. Tech. (2023).

[28] Y. Du, Z. Hua, C. Liu, R. Lv, W. Jia, M. Su, ATR-FTIR combined with machine learning for the fast non-targeted screening of new psychoactive substances, Forensic Sci. Int. 349 (2023), 111761.

[29] H. Zhu, J. Yang, N. Huang, Assessment of the Generalization Abilities of Machine-Learning Scoring Functions for Structure-Based Virtual Screening, J. Chem. Inf. Model. 62 (22) (2022) 5485–5502.

[30] Y. Chen, X. Yu, W. Li, Y. Tang, G. Liu, In silico prediction of hERG blockers using machine learning and deep learning approaches, J. Appl. Toxicol. 43 (10) (2023) 1462–1475.

[31] N. Schaduangrat, N. Anuwongcharoen, P. Charoenkwan, W. Shoombuatong, DeepAR: a novel deep learning-based hybrid framework for the interpretable prediction of androgen receptor antagonists, J Cheminform 15 (1) (2023) 50.

[32] M. Tahir ul Qamar, X.-T. Zhu, L.-L. Chen, L. Alhussain, M.A. Alshiekheid, A. Theyab, M. Algahtani, Target-Specific Machine Learning Scoring Function Improved Structure-Based Virtual Screening Performance for SARS-CoV-2 Drugs Development, Int. J. Mol. Sci. 23 (19) (2022) 11003.

[33] Y. He, G. Liu, S. Hu, X. Wang, J. Jia, H. Zhou, X. Yan, Implementing comprehensive machine learning models of multispecies toxicity assessment to improve regulation of organic compounds, J. Hazard. Mater. 458 (2023), 131942.

[34] Y. Lin, Y. Zhang, D. Wang, B. Yang, Y.Q. Shen, Computer especially AI-assisted drug virtual screening and design in traditional Chinese medicine, Phytomedicine 107 (2022), 154481.

[35] D. Herman, M.M. Kandula, L.G.A. Freitas, C. van Dongen, T. Le Van, N. Mesens, S. Jaensch, E. Gustin, L. Micholt, C.H. Lardeau, C. Varsakelis, J. Reumers, S. Zoffmann, Y. Will, P.J. Peeters, H. Ceulemans, Leveraging Cell Painting Images to Expand the Applicability Domain and Actively Improve Deep Learning Quantitative Structure-Activity Relationship Models, Chem. Res. Toxicol. 36 (7) (2023) 1028–1036.

[36] J. Wang, C. Lou, G. Liu, W. Li, Z. Wu, Y. Tang, Profiling prediction of nuclear receptor modulators with multi-task deep learning methods: toward the virtual screening, Brief Bioinform 23(5) (2022).

[37] M. Kumari, N. Subbarao, Convolutional neural network-based quantitative structure-activity relationship and fingerprint analysis against inhibitors of anthrax lethal factor, Future Med. Chem. 15 (10) (2023) 853–866.

[38] M. Hashemi, A.N. Vattikonda, V. Sip, S. Diaz-Pier, A. Peyser, H. Wang, M. Guye, F. Bartolomei, M.M. Woodman, V.K. Jirsa, On the influence of prior information evaluated by fully Bayesian criteria in a personalized whole-brain model of epilepsy spread, PLoS Comput Biol 17(7) (2021) e1009129.

[39] D.E. Graff, E.I. Shakhnovich, C.W. Coley, Accelerating high-throughput virtual screening through molecular pool-based active learning, Chem. Sci. 12 (22) (2021) 7866–7881.

[40] M. Nasser, N. Salim, H. Hamza, F. Saeed, I. Rabiu, Improved Deep Learning Based Method for Molecular Similarity Searching Using Stack of Deep Belief Networks, Molecules 26 (1) (2020) 128.

[41] Y. Shi, X. Zhang, Y. Yang, T. Cai, C. Peng, L. Wu, L. Zhou, J. Han, M. Ma, W. Zhu, Z. Xu, D3CARP: a comprehensive platform with multiple-conformation based docking, ligand similarity search and deep learning approaches for target prediction and virtual screening, Comput. Biol. Med. 164 (2023), 107283.

[42] Y. Huang, H. Zhang, S. Jiang, D. Yue, X. Lin, J. Zhang, Y.Q. Gao, DSDP: A Blind Docking Strategy Accelerated by GPUs, J. Chem. Inf. Model. 63 (14) (2023) 4355–4363.

[43] M. Riedl, S. Mukherjee, M. Gauthier, Descriptor-Free Deep Learning QSAR Model for the Fraction Unbound in Human Plasma, Mol. Pharm. (2023).

[44] G. Turon, J. Hlozek, J.G. Woodland, A. Kumar, K. Chibale, M. Duran-Frigola, First fully-automated AI/ML virtual screening cascade implemented at a drug discovery centre in Africa, Nat. Commun. 14 (1) (2023) 5736.

[45] T. Li, Z. Liu, S. Thakkar, R. Roberts, W. Tong, DeepAmes: A deep learning-powered Ames test predictive model with potential for regulatory application, Regul. Toxicol. Pharm. 144 (2023), 105486.

[46] A.D. Kalian, E. Benfenati, O.J. Osborne, D. Gott, C. Potter, J.C.M. Dorne, M. Guo, C. Hogstrand, Exploring Dimensionality Reduction Techniques for Deep Learning Driven QSAR Models of Mutagenicity, Toxics 11(7) (2023).

[47] U. Panwar, A. Murali, M.A. Khan, C. Selvaraj, S.K. Singh, Virtual Screening Process: A Guide in Modern Drug Designing, Methods Mol. Biol. 2714 (2024) 21–31.

[48] Y. Yuan, F. Pan, Z. Zhu, Z. Yang, O. Wang, Q. Li, L. Zhao, L. Zhao, Construction of a QSAR Model Based on Flavonoids and Screening of Natural Pancreatic Lipase Inhibitors, Nutrients 15(15) (2023).

[49] W.C. Chou, Q. Chen, L. Yuan, Y.H. Cheng, C. He, N.A. Monteiro-Riviere, J. E. Riviere, Z. Lin, An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice, J. Control. Release 361 (2023) 53–63.

[50] H. Wang, G. Zhu, L.T. Izu, Y. Chen-Izu, N. Ono, M.D. Altaf-Ul-Amin, S. Kanaya, M. Huang, On QSAR-based cardiotoxicity modeling with the expressiveness-enhanced graph learning model and dual-threshold scheme, Front. Physiol. 14 (2023) 1156286.

[51] Y.L. Liu, R. Moretti, Y. Wang, B. Bodenheimer, T. Derr, J. Meiler, Integrating Expert Knowledge with Deep Learning Improves QSAR Models for CADD Modeling, bioRxiv (2023).

[52] M. Dablander, T. Hanser, R. Lambiotte, G.M. Morris, Exploring QSAR models for activity-cliff prediction, J. Cheminform. 15 (1) (2023) 47.

[53] Y. Wu, M. Li, J. Shen, X. Pu, Y. Guo, A consensual machine-learning-assisted QSAR model for effective bioactivity prediction of xanthine oxidase inhibitors using molecular fingerprints, Mol. Divers. (2023).

[54] E.A. Sosnina, S. Sosnin, M.V. Fedorov, Improvement of multi-task learning by data enrichment: application for drug discovery, J. Comput. Aided Mol. Des. 37 (4) (2023) 183–200.

[55] R.D. Shirwaikar, U.D. Acharya, K. Makkithaya, S. Srivastava, U.L. Lewis, Optimizing neural networks for medical data sets: A case study on neonatal apnea prediction, Artif. Intell. Med. 98 (2019) 59–76.

[56] L.K. Tsou, S.H. Yeh, S.H. Ueng, C.P. Chang, J.S. Song, M.H. Wu, H.F. Chang, S. R. Chen, C. Shih, C.T. Chen, Y.Y. Ke, Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery, Sci. Rep. 10 (1) (2020) 16771.

[57] I.I. Baskin, The power of deep learning to ligand-based novel drug discovery, Expert Opin. Drug Discov. 15 (7) (2020) 755–764.

[58] M. Puttagunta, S. Ravi, Medical image analysis based on deep learning approach, Multimed. Tools Appl. 80 (16) (2021) 24365–24398.

[59] H.M. Ashtawy, N.R. Mahapatra, Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment, J. Chem. Inf. Model. 58 (1) (2018) 119–133.

[60] J. Li, A. Fu, L. Zhang, An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking, Interdiscip. Sci. 11 (2) (2019) 320–328.

[61] A. Kensert, J. Alvarsson, U. Norinder, O. Spjuth, Evaluating parameters for ligand-based modeling with random forest on sparse data sets, J Cheminform 10 (1) (2018) 49.

[62] I. Cortes-Ciriano, Benchmarking the Predictive Power of Ligand Efficiency Indices in QSAR, J. Chem. Inf. Model. 56 (8) (2016) 1576–1587.

[63] V. Mani, M. Prakash, W.C. Lai, Cloud-based blockchain technology to identify counterfeits, J Cloud Comput (heidelb) 11 (1) (2022) 67.

[64] T.K. Mackey, A.J. Calac, B.S. Chenna Keshava, J. Yracheta, K.S. Tsosie, K. Fox, Establishing a blockchain-enabled Indigenous data sovereignty framework for genomic data, Cell 185 (15) (2022) 2626–2631.

[65] S.J. Trenfield, A. Awad, L.E. McCoubrey, M. Elbadawi, A. Goyanes, S. Gaisford, A. W. Basit, Advancing pharmacy and healthcare with virtual digital technologies, Adv. Drug Deliv. Rev. 182 (2022), 114098.

[66] P.E. Velmovitsky, F.M. Bublitz, L.X. Fadrique, P.P. Morita, Blockchain Applications in Health Care and Public Health: Increased Transparency, JMIR Med. Inform. 9 (6) (2021) e20713.

[67] R.W. Seaberg, T.R. Seaberg, D.C. Seaberg, Use of Blockchain Technology for Electronic Prescriptions, Blockchain Healthc Today 4 (2021).

[68] M. Elbadawi, L.E. McCoubrey, F.K.H. Gavins, J.J. Ong, A. Goyanes, S. Gaisford, A. W. Basit, Harnessing artificial intelligence for the next generation of 3D printed medicines, Adv. Drug Deliv. Rev. 175 (2021), 113805.

[69] G. Gursoy, C.M. Brannon, M. Gerstein, Using Ethereum blockchain to store and query pharmacogenomics data via smart contracts, BMC Med. Genomics 13 (1) (2020) 74.

[70] M. Raghavendra, Can Blockchain technologies help tackle the opioid epidemic: A Narrative Review, Pain Med. 20 (10) (2019) 1884–1889.