Bilad Alrafidain Journal for Engineering Science and Technology
https://dx.doi.org/10.56990/bajest/2025.040207
ISSN: 2073-9524
Pages:67-85

Advanced Deep Learning for Accelerated Drug Discovery: Approaches, Challenges, and Future Expectations

Marwa Abdulkareem Dawoud[1*], Jumana Waleed[1]
[1]Department of Computer Science, College of Science, University of Diyala, Diyala, 32001, Iraq
scicompms232405@uodiyala.edu.iq[1], jumanawaleed@uodiyala.edu.iq[2]

Abstract
Advanced models of deep learning have been transformational tools for discovering drugs and solving problems related to cost, time, and complexity. Using complex network frameworks such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), generative adversarial networks (GANs), and graph neural networks (GNNs), researchers have made major improvements in predicting drug-target interactions, performing virtual screens, and developing novel drugs. Those frameworks effectively occupy elaborate biochemical relations and precisely imitate complicated molecular reciprocities. Nevertheless, significant issues remain, like data shortages arising from restricted access to (high-quality) datasets, model predictions' interpretability, and the scalability to include considerable and assorted datasets. To efficiently address these issues, innovative strategies, including diverse techniques of data augmentation, like molecular graph transformations, have been applied to improve datasets. Reinforcement learning has helped improve molecular structures to accomplish desired characteristics, while ensemble learning, which integrates various model structures, has proven effective in improving prediction reliability. Incorporating multi-modal datasets, like pharmacophores properties, 3D molecular representations, and molecular graphs, increases the accuracy of prediction by occupying spatial and even functional molecular properties. Despite these advances, issues remain in multi-drug modeling, drug resistance management, and accurate toxicity prediction. Future works focus on the importance of explainable AI in strengthening model interpretability, with hybrid structures that incorporate machine learning and experimental feedback to simplify the therapeutic scheme. By addressing these challenges and adopting innovative approaches, deep learning is set to revolutionize drug discovery, enabling a more efficient, accurate, and reliable development pipeline for novel therapeutics. This study highlights how model interpretability and confidence can be enhanced by integrating multigene data and leveraging explainable AI techniques. By focusing on these cutting-edge developments, this study aims to provide practical insights for researchers and practitioners to accelerate the development of safe, effective, and personalized therapeutics.

Keywords: Advanced Deep learning, Drug discovery, Multi-omics data, Drug design, Virtual screening

Article history: **Received:** 16 Jan 2025**, Accepted:** 4 May 25**, Published:** 15 Sep 2025.

* Corresponding author:scicompms232405@uodiyala.edu.iq

ISSN: 2073-9524
Pages:67-85

1. Introduction

In the past, medicinal chemists operating in labs were primarily responsible for drug discovery and development, which involved extensive testing, validation, and synthetic processes. This approach required a large amount of time and financial commitments to get a single medicine to clinical trials. Despite this, the proliferation of multi-omics data and advances in computational technology have produced a plethora of tools in cheminformatics, pharmacology, and bioinformatics. Drug development has been greatly advanced by these developments. The advent of Artificial Intelligence (AI) (including Deep and Machine Learning (DL& ML)) has further changed conventional drug development techniques. Large biological datasets that are spread across international databases are becoming useful tools for ML and DL techniques. The time, labor, and financial resources required for drug development are reduced when these technologies are used to expedite the process of finding therapeutic compounds [1].

DL is shedding new light on drug discovery. While the reliance on the availability of large training datasets is still an 'Achilles' heel' of DL in this domain, recent advancements have proven to be quite successful in applying neural networks to low-data scenarios. Increasingly, such research demonstrates successful approaches to constructing DL models that perform well under data scarcity frameworks. Despite advancements, there are deeper challenges that still remain under the data. However, the advent of DL in drug discovery has brought about more accurate and insight data-driven, along with the lower time and cost that a drug is usually associated with in the development process [2]. In a much different fashion to traditional computational approaches, deep learning is able to synthesize and learn through large complex data sets of layer upon layer neural networks able to recognize patterns within biological data, including genomic, proteomic, and metabolomics data [3].

This ability will prove essential in drug discovery where DL models have depicted great guarantee in new target identification, predicting drug-target interaction, drug metabolism and toxicity modeling, and even de novo drug-like molecule engines. Convolutional and graph neural networks (CNNs & GNNs) enabled researchers to assess analytical models of the structure and function of the probed drug compounds with exceptional precision, in some cases, greatly surpassing traditional machine learning. DL is now extensively used in virtual screening, (absorption, distribution, metabolism, excretion, & toxicity (ADMET)), prognosis, protein structure modeling, and even designing clinical trials, contributing significantly at each drug discovery level, starting from fundamental and then going to preclinical and clinical phases. As DL models continue to evolve, their applications are anticipated to advance and optimize pipelines of drug discovery, providing the pharmaceutical industry with more effective pathways for evolving safe and innovative therapeutics [2]. The main contribution of this investigation:

1. Explores the transformative role of DL in drug discovery, leveraging advanced architectures like GNNs and Generative Adversarial Networks (GANs), along with innovative Transfer Learning or Multi-Task Learning approaches to address the scarcity of data.
2. It emphasizes multi-omics data integration and explainable AI to enhance predictive accuracy and model interpretability, showcasing practical applications in de novo drug design, toxicity prediction, and ADMET property assessment.
3. The study discusses key challenges, including noisy data, high computational demands, and ethical concerns. It proposes solutions such as data augmentation and ethical frameworks and forecasts future innovations to accelerate the development of effective and safe therapeutics.

2. Deep Learning based Drug Discovery

The availability of adequate data has effectively transformed AI techniques into improved machine learning techniques in recent decades to address fundamental issues. One of the greatest ways to address problems involving various variables and large amounts of data is through ML. DL approaches represent resilient and effective tools for processing the vast amounts of information

ISSN: 2073-9524
Pages:67-85

generated by the sheer volume of data being generated in various sectors, and have recently replaced ML techniques. A subfield of ML called "DL" was created to process extremely complicated data and make judgments based on the analysis [4]. As Fig. 1 illustrates, AI has occasionally been used in a variety of pharmaceutical and healthcare fields [5]. DL methods have been integrated into state-of-the-art drug discovery to efficiently process the large volumes of data created in this field, driving advancements in research and development. Numerous models of ML and DL-based applications have been utilized in drug discovery and development. StackCBPred, LigGrep, LS-align, TrixX, and DrugFinder are a few prominent ML-based models. When compared to ML techniques, DL models have provided superior performance. As a result, they have lately become particularly promising instruments in drug discovery research. DeepDTA, WideDTA, PADME, DeepAffinity, and DTI-CNN are a few notable DL-based models. Drug development and discovery are greatly advanced by these models [5].

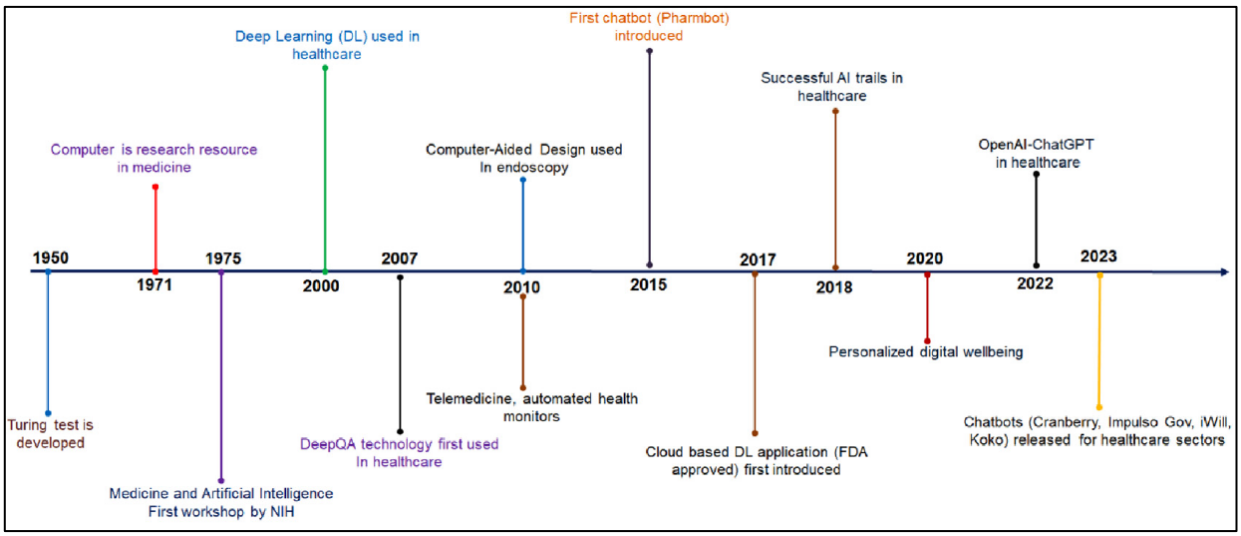

Fig. 1 The turning point of AI-related accomplishments in the healthcare and pharmaceutical areas [5]**.**

2.1 Virtual Screening

Another computer-based method in drug development is virtual screening, which looks for basic libraries of small compounds to identify hits that have the best chance of attaching to a drug destination. The chemical compound's biological, topological, physical, and chemical characteristics and targets are used in virtual screening. The techniques of virtual screening can be categorized into two structures. The first is structure-based, which models and conceptualizes interactions using the three-dimensional structure of chemical and targets compounds.

The three-dimensional structure can be obtained via nuclear magnetic resonance or crystallography (X-ray). Docking could be utilized to specify how a molecule interacts with a particular target once the three-dimensional structural data has been put together. The second category is nanostructure-based and can be further classified into two groups: modeling of protein-chemical, which involves integrating non-structural headlines with targets (at the input-level), and ligand-based virtual screening, which models and analyzes interactions with targets using the molecular properties of compounds [6]. Numerous surveys have demonstrated that DL outperformed other ML models in terms of virtual screening, particularly owing to their significant implementation in de novo molecular design, which utilizes data sequence to generate molecules of desired properties. DL-based tools that are utilized in virtual screening [7].

2.2 De Novo Drug Design

The process of De novo drug design works on creating drug-like molecules using computational

ISSN: 2073-9524
Pages:67-85

methods, starting from scratch and not depending on a predefined structure. The rise in artificial intelligence techniques has accelerated drug discovery and opened up new possibilities for novel drug schemes, such as the formation of new molecular structures from atomic building blocks without prior relationships. But, there are some significant variations between de novo drug design and traditional one. In traditional drug design, a structure-based procedure was considered, which relies on the properties of the protein binding site of the biological target. AI is a growing field that has affected drug discovery.

Thus, AI has also had an impact on the innovative medication design strategy. It uses knowledge to generate new chemical entities, such as ligands and receptors. The biological targets in this case are referred to as receptors or their active ligands, respectively. Modeling the receptor's active site or ligand pharmacophore is essential to creating a molecule in de novo drug creation. For the design of innovative drugs, a number of models based on ML and DL models have been proposed. Different DL-based models have been created on occasion over the years to help in the development of novel medications.

Another recently noted model for novel drug molecule design is druGAN. A deep generative autoencoder (AAE) model has been implemented to develop new molecules with anticancer properties. In the same manner, incorporating reinforcement learning with a hybrid VAE resulted in creating a novel drug design model known as PaccMannRL. This model uses genomic transcription data to efficiently design anticancer molecules. More recently, a new drug design model that utilizes GNNs was introduced. This Model is able to effectively create indispensable novel quinolone compounds, assess their accessibility, toxicity, and pharmacokinetics, and their potential as drugs. This illustrates the capacity and depth of impact ML/DL algorithms can have with respect to drug designing [5]. From time to time, more recent models to drug design have creatively incorporated evolutionary algorithms:

A. Using generative models for creating new molecular structures offers a novel approach due to their ability to utilize synthetic heuristics or priors, effectively replacing the search-through molecular space paradigm with simply validating promising results. These models do especially well when learning from existing datasets of molecules and creating new molecules that share the same traits. The molecules produced through these methods could then be filtered by a number of criteria, such as ease of synthesis, affinity towards selected proteins, biological activity, and other relevant physicochemical parameters. In recent years, several generative models have been proposed for drug design employing various representations of molecules, including, but not limited to, SMILES strings, 2D or 3D molecular graphs, Morgan fingerprints, and even images of the molecules themselves. Moreover, whether they do or do not identify protein targets makes available a rich collection of models and methodologies intended for various data formats and conditions. Many of these approaches draw inspiration from language models and generative pre-trained transformers (which known with GPT), further enhancing their ability to generate novel drug-like compounds efficiently [8].

B. Reinforcement learning for drug optimization: It is a sub-section of ML that combines ANNs with the architectures of reinforcement learning. This combination has been effectively utilized to produce new drug design methods utilizing various Neural Networks, encompassing RNNs, CNNs, competitive generative networks, and auto-encoders [9].

2.3 Toxicity Prediction

Drug toxicity refers to the prediction of unwanted or harmful properties in drug-like molecules, a critical factor contributing to the high costs of drug development. Since toxicity is directly linked to drug safety, accurately predicting side effects and assessing safety are essential components of the drug development process. However, conducting laboratory-based toxicity studies during drug

ISSN: 2073-9524
Pages:67-85

development is time-consuming and resource-intensive. To address this challenge, computational models have been developed to reduce both the time and cost associated with toxicity prediction. Recently, a model called DeepTox, based on a three-layer deep neural network (DNN), has been introduced to predict the toxicity of drug-like molecules or compounds. This model utilizes molecular descriptors ranging from 0D to 3D as input for DNNs. Additionally, a deep learning-based toxicity prediction model named Deep-PK has been developed to evaluate the toxicity and pharmacokinetics of small molecules. These developments emphasize the ultimate importance of ML/DL models in the drug toxicity prediction process, making it more efficient and cost-effective [5].

A. Deep learning Models to Predict ADMET-properties: The ADME Toxicity (ADMET) issues the drug- discovery and development process are faced with have been vaguely referred to as the absorption, distribution, metabolism, excretion and adverse toxicity properties which are major known causes of failure that leads several molecules dissolution from the drug development pipeline and subsequently initiates an increase in wastage of valuable time and resource. This surging interest has changed the focus to earlier stages of prediction of ADMET properties of candidates, aiming to increase the probability of success for the compounds proceeding to the further stages of the drug development process. AI technology has successfully been used to create models and tools for the prediction of ADMET, resulting in the simplification of the evaluation of drug-like molecules prior to rigorous testing. Also, companies which take up clinical studies have witnessed enhanced success rates with refinement of their research approaches using the strategies which are AI based. The changes have resulted in the improvement of the preclinical and clinical studies, which proves the use of AI in drug development and discovery to be promising and benchmarking. This is why AI is advocated in the field of modern medicine [1].

B. Case Studies on Toxicity and Safety Assessments: Drug-induced toxicity poses a major issue to late-stage drug attrition. The safety profile of a potential drug candidate must be closely monitored to neglect troubled drug interactions. Most pharmaceutical corporations and regulatory bodies in numerous countries (as of 2020) have relied on in vivo and vitro testing. While this approach supplies scientific researchers with significant data, it has some drawbacks. Some toxicity patterns remain undiscovered at the preclinical level owing to metabolic and physiological discrepancies between people and others, and synergistic effects. Sitaxentan is an instance of a drug that is not hepatotoxic in animal studies but causes significant liver injury in humans

2.4 Protein-Ligand Interaction Prediction

Conformational changes in target proteins triggered by ligand binding influence various biological processes. Modern medicine has significantly advanced by leveraging the capability of modifying the structure and function of proteins throughout the use of small molecules as therapeutic agents for treating diseases. Although protein-ligand interactions (PLI) are important in medicine and biology, and insight into the multiple factors governing ligand recognition, involving hydrophobicity, hydrogen bonding, and p-interactions, developing and validating accurate predictive PLI models for drug discovery continues to be a complex challenge.

Relying solely on experimental methods to identify and confirm protein-ligand interactions can be time-consuming and costly. On the other hand, computational approaches offer a more efficient alternative by screening vast libraries of compounds and narrowing them down to a smaller selection of ligands with a higher likelihood of binding to the target protein. By doing so, accurate PLI prediction algorithms can significantly speed up the discovery of new therapeutics, help eliminate potentially toxic drug candidates, and provide valuable guidance in medicinal chemistry [10].

ISSN: 2073-9524
Pages:67-85

2.5 Drug Repurposing

Due to the rate of high attrition, the impediments of large costs, a slow march of drug discovery and development, the fact that repurposing of the "old" drugs becomes more and more popular the treatment of not only usual but also very rare diseases, since it comprises the application of lower-risk compounds, with possibly lesser development costs, and shorter development periods, it is definitely worthwhile one. Several data-driven experimental approaches have been proposed to identify reusable drug candidates; however, there are also significant technological and regulatory challenges that need to be addressed [11].

Drug repurposing offers a cost-effective strategy for finding new medical applications for already approved drugs. Advances in AI have enabled the systematic identification of potential repurposing opportunities by leveraging large-scale datasets, accelerating drug development, and minimizing risks through computational analysis. This study focuses on supervised machine learning (ML) approaches that utilize publicly available databases and existing knowledge resources. While most applications have been in anticancer drug therapies, the methodologies explored are broadly applicable to other conditions, including COVID-19 treatment.

A key focus is placed on comprehensive target activity profiles, which facilitate systematic drug repurposing by expanding the known target spectrum to include potent off-target effects with therapeutic potential for new indications. However, the limited availability of clinical patient data and the prevailing emphasis on genetic aberrations as primary drug targets may restrict the effectiveness of anticancer drug repurposing strategies that rely solely on genomic-based insights. Functional testing of cancer patient cells exposed to various targeted therapies and drug combinations offers an additional layer of valuable, real-world data that can enhance AI-driven, tissue-specific drug repurposing approaches [12].

3. Datasets in DL-based Drug Discovery

Drug discovery faces significant challenges related to data limitations and quality. The small and often noisy datasets used in this domain can restrict model performance due to insufficient diversity and volume, compounded by inadequate negative examples (inactive molecules). The studies of Maharana et al. [13], Bhati et al. [14] and Lavecchia [15], suggest introducing bias and hindering reliable learning. Integrating data from diverse sources, such as high-throughput screenings, literature, and databases, further complicates pre-processing due to inconsistencies and heterogeneity in data types and formats [15]. Techniques like QSAR (Quantitative Structure-Activity Relationship) models frequently contend with sparse and highly correlated feature sets, necessitating careful handling to avoid misleading predictions [16] and [17].

Additionally, representing molecules and proteins in computer-readable formats such as SMILES strings or molecular graphs often results in the loss of critical 3D conformational and dynamic properties [17]. Encoding biological data effectively for deep learning models remains a significant technical challenge, as does the extensive effort required in pre-processing steps like molecular structure standardization, deduplication, and inconsistency correction. Pre-processing is an important procedure because it can have a positive effect on the results of the experiments and the reliability of the data [18]. It is also a known fact that selection bias is common in the dataset aeration process, and this error in the data may lead to overfitting, which may seem like the models are working well with he given data, but in reality, they do not perform well on the unseen data. Bias control and data balancing model construction (that may be used to fight and eliminate these factors) are both important parts in the process of developing effective and reliable drug discovering systems [19], [20]. Table 1 illustrates the datasets utilized for drug discovery.

ISSN: 2073-9524
Pages:67-85

Table 1: Datasets used (Public & Proprietary).

Datasets	Link to the dataset	Description	Source/Features	Last Update
Binding DB	https://www.bindingdb.org/bind/	A publicly accessible library of measured binding affinities that focuses on interactions between tiny, drug-like compounds and proteins thought to be therapeutic targets	Contains curated binding affinity data for over 20,000 proteins and drug-like compounds; supports drug discovery research	2023-08-01
ChEMBL	https://www.ebi.ac.uk/chembl/	European Bioinformatics Institute, known as EBI, supports a hand-selected chemical database of bioactive compounds with drug-like characteristics.	Includes data on bioactive compounds, their targets, and associated activities; useful for cheminformatics and pharmacology	2023-07-15
Bioassay Datasets from PubChem	https://archive.ics.uci.edu/	A repository of bioassay data, supplying information on the biological activities of small molecules	Comprises millions of bioassay records, integrating high-throughput screening results; supports predictive modeling in drug discovery	2023-06-20
DrugBank	https://go.drugbank.com/	A thorough resource that blends comprehensive medication target knowledge with detailed drug data	Offers information on approved drugs, investigational drugs, and their interactions with biological targets	2023-08-25
Drug Target Commons (DTC)	drugtargetcommons.fimm.fi	A community-driven platform for sharing and analyzing drug-target interaction data	Includes crowd-sourced annotations, enabling systematic analysis of drug-target relationships	2023-07-30
Broad Bioimage Benchmark Collection	https://bbbc.broadinstitute.org/	These datasets are often available through specific research institutions or collaborative projects	Features a wide variety of imaging datasets, including Cell Painting assay data, from collaborative projects	2023-08-10
High-throughput Screening (HTS)	-	HTS data can be accessed through platforms like PubChem BioAssay.	Provides data on millions of compounds tested against various targets	2023-06-05
Preclinical Antibody Data	http://i.uestc.edu.cn/DOTAD/	Databases like the Antibiotype provide information on antibodies, including preclinical data	Includes preclinical testing and validation results for various antibodies	2023-07-01
PubChem	https://pubchem.ncbi.nlm.nih.gov/	Accessible chemistry database supported by NCBI, which is the National Center for Biotechnology Information; biological activities of small molecules	Integrates chemical structure, bioactivity, and assay data for over 200 million compounds; supports cheminformatics research	2023-09-01

4. Challenges and Limitations

4.1 Scarcity and Quality of Data

Another vital point is that drug discovery data are available in the highest amount, however, the quality of the data is still a lot lower compared to other ML areas. That said it also goes with the territory that health data are very sensitive and often not of the best quality. Furthermore, developments made in hardware have prompted DL revolution, but the field of drug discovery is still not sufficiently developed to the point of machine learning field where training a model requires a lot of computing resources.

Drug discovery is a very challenging process where the one is tasked with predicting the interactions of the molecular structures and understanding their characteristics. A model is supposed to be able to handle various sizes of molecular structures but it has to adapt to different datasets and make sure that the findings are largely applicable. The models can be used to generate

ISSN: 2073-9524
Pages:67-85

novel molecular structures, however, they can also fail to solve the chemical validity, novelty, and synthesis problem. Furthermore, the models might just be useless and the new molecules need to be taken with caution. The same applies to the evaluation of these models as there is no specific benchmark that can be used to compare their performance.

Current datasets could be split or with prejudices in them, so it is quite difficult to exactly compare various approaches and test if the model performs well [16]. One of the problems with the use of DL in drug discovery is the existence of several obstacles and the issue of understanding the data. The concept of data scarcity concerns that the type of data needed is not enough, is one of the key problems. The lack of a larger number of data points to choose from can result in the misuse and poor generality of the training datasets, which will bring biases and worsen generalization. Furthermore, very often comprehending the decisions of such models is a tough job, as these models work as "black boxes." Despite these limitations, approaches like data augmentation, transfer learning, and self-supervised learning are utilized that enhance the model performance [13]. Another example is where the interpretability of SHAP, LIME, and attention mechanisms is a tool for clarification, while the models that are hybrids and encompass different features merge for the production of accurate models [18].

The essential noise in chemical and biological datasets is one of the toughest problems to deal with for ML models because it is always present and adversely affects the reliability of ML models. This noise is frequently attributable to other factors , for instance, the high complexity of biochemical assays or variation in conditions when experiments were conducted. Also, the lack of negative samples or inactive compounds makes model training a harder task since publication bias is common. It means that the studies are likely to show only a positive or a successful result. That results in datasets where there is no representation of the true distribution of active and inactive compounds. Imbalanced data can, as a result, make it difficult for models to generalize effectively, as these might show good results only on the training data and fail on new and unseen data. Consequently, noisy and skewed datasets can cause models to overfit or develop biases toward specific patterns, affecting their ability to make accurate predictions for diverse drug candidates [16].

4.2 DL Model Interpretability

The field of drug development faces several challenges; The main difficulty is the attainability of high-quality datasets fit for training AI models. Although the availability of chemical and biological data is expanding, data quality often remains suboptimal, making effective data processing essential. Accordingly, retrieving data from databases can be costly, adding to the overall expenses of drug development. That said, high-quality datasets do exist within pharmacology, and fostering collaboration between technology and pharmaceutical companies could accelerate drug discovery and development as displayed in Fig. 2. Such datasets can help in training and testing AI models, which may solve the problem of data availability. Another major challenge is the ability to interpret and understand AI model predictions. Deep learning models involve a large number of complex parameters and layers, making it difficult for non-experts to understand the decision-making processes in the context of drug discovery and development. As a result, improving the interpretability of deep learning models is essential for advancing drug research, but achieving this remains a significant hurdle. [5].

4.3 Balancing ML and DL in Drug Discovery

Machine learning (ML) is used in drug discovery when limited data is available, as it relies on manual feature extraction and offers ease of interpretation and efficiency with small datasets. On the flip side, deep learning (DL) is very good in analyzing large and complex datasets like molecules, but it needs high computation power and often is like a "black box." DL is the first choice when a great accuracy is required with big data sets, and ML is chosen for jobs that need to explainability. The most effective way to keep the balance performance and the explanation is to use both these methods in the same models. It is undeniable that the future of research is

ISSN: 2073-9524
Pages:67-85

centred around growing integrated ML and the DL track to raise the drug discovery efficiency bar [15].

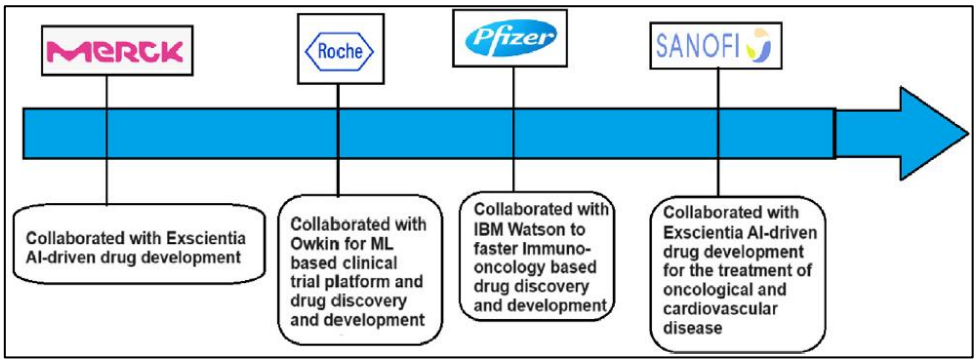

Fig. 2 Collaboration between pharmaceutical and technology companies to develop and discover AI-based drugs [5].

4.4 Domain-Specific Integration

Integration of multiple types of data, such as molecular structures, protein–protein interactions and biological pathways, has enabled building DL models able to discern the complex dependencies/relationships that are important for potentiation of the action of a drug [9], [21]. The use of sophisticated DL models has led to multi-dimensional data being used, such as 3D Molecular representations, and graph-based models can bring high-level biological and chemical properties together for predictions [21], [22]. DL models, such as RNNs and CNN, are generalized to incorporate pharmacological knowledge [23], [24]. Yet, with advancement, the challenges persist, such as there not currently being a best quality solution or the insufficiency of even the most complete data-sets and similarly, how to model even more complex biological processes without over-generalization. This scheme we developed will help the DL models more compliance with the biological truth and have more reliable and accurate predictions during drug discovery [24].

4.5 Computational Complexity

The techniques of DL, including RNNs, CNNs, GANs, and AEs, have proven to be highly effective in drug discovery. However, these models are computationally expensive, leading to difficulties in environments with limited access to high-performance computing resources [9].

Training complex neural networks (especially neural networks on graphs) requires more resources of computations. This is because of the size of the parameters involved and also the iterative process of training the models. Graph convolutional and attention networks are complex to work with and can handle large datasets, but are expensive in terms of CPU and memory. In addition, the performance of these models is often reliant on the scale of the datasets. As datasets scale up, maintaining a model without a dramatic increase in computation time can be challenging. Hybrid models (from molecular graphs with augmented fingerprint features) can improve predictive abilities but also increase computational costs. However, optimizing such hybrid models requires careful tuning of hyper parameters to balance computational feasibility and model performance [25].

4.6 Ethical and Legal Considerations in AI-based Drug Discovery

AI in pharma manufacturing needs to be responsible and transparent. Bias must be mitigated to promote fair treatment, but, importantly, also to comply with data protection legislation such as the GDPR and HIPAA. Moreover, obtaining regulation approval from agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) is essential for creating clear legal responsibility when medical accidents occur. Intellectual property rights also raise concerns regarding ownership of AI-generated inventions.

The promise of AI in revolutionizing the drug discovery and improving predictive accuracy can only be fulfiled but for its safe and equitable use, robust ethical and legal frameworks have to be put

ISSN: 2073-9524
Pages:67-85

in place. This includes the definition of clear guidelines for AI model training, the enablement of transparency in AI-driven decisions, safeguards protecting patient data and definitions about those who are legally liable in the case that _ orexcursus [58] These could manifest themselves as legal standpoint.

5. Recent Innovations and Emerging Trends

5.1 Advances in Molecular Representation Using GNNs

Graph Neural Networks (GNNs) have exhibited remarkable potential in drug discovery, in particular for data represented by relational graphs, such as chemical compound structures, drug-target interactions (DTIs), protein-protein interactions (PPIs), and patient-disease associations. The flagship types of GNNs are Graph Convolution Networks (GCNs) that employ convolutions as well as Graph Attention Networks (GATs) using attention mechanisms to achieve more localised learning. GNNs have been utilized in a wide variety of applications: Predicting the chemical and biological properties of compounds by modeling the molecular graph structural relationships, predicting the drug-target interaction, and designing de novo drugs. to facilitate the generation and optimization of novel chemical structures based on learn adaptation model.The ability of GNNs to perform structured learning has further promoted their application across various fields [22].

GNNs have the attractive property of capturing complex graph structures and their relationships, effectively enabling better learning of more informative representations, leading to improved prediction performance for drug discovery. Some differences with these networks reside in the robustness and training complexity. However, they still present computational complexity and face some challenges, such as high computing requirement; or difficulty the training large/complex graphed problems; generalization issues; compatibility with other systems [22].

5.2 Drug Design and Virtual Screening Using GANs

In drug research, Generative Adversarial Networks (GANs) have shown great potential, particularly in enhancing virtual screening and developing drug discovery. By learning from existing data, GANs can generate high-quality chemical structures with certain desired properties, including biological activity against a target. These models help identify promising therapeutic leads and streamline the drug development process by generating candidate molecules that meet predefined criteria [9], [14]. In addition, GANs have the potential to improve virtual screening by constructing molecular architectures that are more likely to bind to the desired target proteins, thereby accelerating drug discovery through faster identification of potential drug candidates [26]. Moreover, GANs enable the exploration of enormous chemical spaces by creating novel compounds that may not exist in traditional databases, improving the diversity of drug candidates and improving the candidate drugs [27]. GANs were used in the DeepCancerMap platform to evaluate the effectiveness of anticancer drugs, streamlining high-volume virtual screening and drug repurposing [28].

5.3 Multi-task and Few-shot Learning in Drug Discovery

Sharing the knowledge between different tasks can lead to information crossing between different tasks to enhance the performance. In drug discovery, tasks often include predicting multiple bioassays, predicting biochemical activities of associated compounds, or countless other multifarious molecular properties [27]. In scenarios where there is limited data, as shared information between tasks can lead to reduced overfitting, MTL aids in information retention. MTL aids in drug discovery bioassays where MTL bioassay outcome predictors outperform single-task models [17]. MTL aids in the modeling of structure-activity relationships (SAR) to biological targets, which improves the generalizability of the models predicting the acute biological activities of multiple molecular properties, such as efficacy, safety, and solubility [29].The advantages of MTL include leveraging

ISSN: 2073-9524
Pages:67-85

data from related tasks to improve performance on individual tasks, reducing computational requirements by training a unified model, and effectively addressing data sparsity by sharing information across tasks [17].

Few-shot Learning (FSL) trains models to perform well with minimal data, making it crucial in drug discovery where data for rare diseases or novel compounds is limited [27] [29]. Applications in drug discovery include predicting activity on rare targets with scarce experimental data by leveraging prior knowledge from similar tasks or datasets, identifying potential drug candidates with limited screening data using techniques like meta-learning, and enhancing FSL models through data augmentation techniques, including synthetic molecule generation via generative models like GANs [29]. The advantages of FSL include reducing dependency on large datasets, making it ideal for exploring niche chemical spaces or orphan targets and facilitating rapid hypothesis generation for experimental validation [30].

5.4 Recent Works

Overcoming challenges in drug discovery through deep learning requires a multifaceted approach that addresses key obstacles while leveraging emerging opportunities. Several works utilized DL for drug discovery and tried to overcome most challenges, as demonstrated in Table 2.

Table 2: Developed works in drug discovery.

References	Techniques	Datasets	Advantage	Disadvantage
Chakraborty et al. [5]	CNNs, RNNs, Transformers, GNNs; Reinforcement Learning	ChEMBL, PubChem, DrugBank, SIDER, ChemDB, DTC, STITCH	Significant reduction in drug discovery time and cost; improved accuracy in target identification, DTI, and ADMET predictions.	Requires large and well-curated datasets; computationally expensive for large-scale implementations.
Wu et al. [6]	CNN, RNN, GNN, Transformer, Attention Mechanisms, Pre-trained Models	DUD-E, MUV, BindingDB, DrugBank, Davis, KIBA, PDBBind, UniProt, PubChem	Enhanced prediction accuracy for drug-target interactions and binding affinity; improved interpretability with attention mechanisms.	Computationally intensive; requires high-quality labeled datasets; interpretability challenges for some deep models.
Mouchlis et al. [9]	Deep reinforcement learning, RNNs, CNNs, GANs, Auto-encoders	- Explores broader chemical space; innovative generative design methods - Various de novo drug design datasets involving chemical structures and biological activity	- Synthetic accessibility of designed compounds - Synthetic accessibility of molecules - Advances drug discovery process; enhances novel molecular structure generation; utilizes AI to streamline design	Complex algorithms require significant computational power; models may overfit with limited data
Maharana et al. [13]	Data pre-processing, classification, clustering, data augmentation techniques (flipping, rotating)	Various machine learning datasets from real-world applications	Improves data quality for machine learning; enhances model accuracy by augmenting data	Potential for overfitting; risk of distorting original data during augmentation
Kwon et al. [17]	Ensemble learning (random forest	19 bioassay datasets with bioactivity and	Improved prediction accuracy by combining	Limited generalizability beyond specific bioassay

ISSN: 2073-9524
Pages:67-85

	(RF), meta-learning)	chemical properties	diversified models; enhanced reliability	datasets; increased complexity in meta-learning approach
Zhou et al. [18]	DL, CNNs, RNNs, GNNs, Transformers, VAEs, GANs	Observed Antibody Space (OAS), Sab Dab, custom datasets	Improved efficiency and accuracy in antibody encoding, discovery, optimization, and humanization.	Requires large annotated datasets, high computational costs, and careful preprocessing for specific antibody features.
Lavecchia, A. [19]	Deep Attention Networks (Transformers, GATs, BERT, GPT, BART), Generative Models	ChEMBL, PubChem, BBBC datasets	Improved prediction accuracy in molecular property prediction, drug-target interactions, and de novo drug design.	High computational cost; limited by data availability and quality; challenges in model interpretability.
Kim et al. [22]	DNNs, DTI prediction, de novo drug design	Drug-target interaction and de novo drug design benchmark datasets	Reduces time and cost in drug discovery; predictive for complex drug-target interactions	High complexity in data representation; still evolving techniques, requiring better model generalization
Elbadawi et al. [23]	ML, DL, Reinforcement Learning, NLP	ChEMBL, DrugBank, PubChem, ZINC	Enabled efficient virtual screening, drug repurposing, and de novo drug design; improved prediction accuracy for drug interactions	Computationally intensive; requires high-quality datasets; limited interpretability of some deep learning models.
Lin et al. [25]	RF, Boost, Graph-based DL Models (GCN, GAT, MPNN, Attentive FP), Co-representations DL (FP-GNN, HiGNN, FG-BERT)	- ChEMBL Database - PubChem Database - MMV St. Jude Dataset - Harvard Liver Stage Dataset	- Improved prediction accuracy with FP-GNN and HiGNN - Broad coverage of Plasmodium life cycle stages - multi-stage prediction capabilities	- Imbalanced datasets challenge generalization - High computational demands for graph-based models on large datasets
Wu et al. [28]	DL (FP-GNN architecture), GNN, Fingerprint Networks	ChEMBL Database, NCI-60 Panel, Cancer Cell Lines	High accuracy in predicting anticancer activity (AUC values > 0.9); robust for both target-based and cell-based predictions.	Computationally expensive; dependent on high-quality annotated datasets; less effective for rare or novel targets.
Chen and Gilson [31]	Relational database management, XML integration	Binding DB (biomolecular binding affinity data)	Publicly accessible, supports various query methods, promotes direct deposition of binding data	Limited to biomolecular binding data, dependent on data deposition by experimentalists
Tiqing Liu 2006 [32]	Relational Database for Protein-Ligand Binding	Binding DB	Public access, supports diverse queries	Limited to proteins with available structural data
Yanli Wang [33]	HTS, BioAssay Curation	PubChem BioAssay Database	Largest public HTS data repository, promotes open access	Inconsistent data quality across studies

ISSN: 2073-9524
Pages:67-85

Lenselink [34]	DNNs, QSAR, PCM	ChEMBL Dataset	Superior predictive performance, scalable modeling	Requires extensive data preprocessing
Chen et al., [35]	Variation Auto-encoders (VAEs), CNNs	ChEMBL Dataset	- Ability to Handle Complex Datasets - Deep learning models can analyze large-scale datasets (e.g., ZINC, ChEMBL) with millions of compounds, enabling faster and more comprehensive screening	DL models require large, high-quality, annotated datasets to achieve optimal performance, which can be challenging to obtain, especially in niche areas like rare diseases
Zhavoronkov et al. [36]	Generative Tensorial Reinforcement Learning	DDR1 kinase inhibition assay data	Accelerates drug discovery, identifies novel compounds	Limited to specific protein targets, requires domain-specific dataset
Shtar et al. [37]	Artificial Neural Networks, Graph Similarity	DrugBank Dataset	High accuracy in DDI prediction, scalable	The methods rely heavily on known drug-drug interaction (DDI) networks, which may not be helpful for new drugs or interactions not yet documented in databases like DrugBank
Mendez et al., [38]	Direct Bioassay Data Deposition	ChEMBL Database	Streamlines bioassay data sharing, reduces errors	Dependent on depositor's accuracy
Kim et al., [39]	Bioactivity and Spectral Data Integration	PubChem Database	Increased spectral data coverage, supports diverse studies	Potential for outdated records
Jacquemard and Rognan [40]	ML, Virtual Screening	LIT-PCBA Dataset	Reduces biases, facilitates rigorous benchmarking	Limited to curated target sets
Bento et al. [41]	RDKit-based chemical curation	ChEMBL Database	Automates standardization, improves data reliability	Dependent on predefined curation rules
Kim et al. [42]	Data Integration and Web Interface Enhancements	PubChem Database	Expanded dataset, improved accessibility	Requires frequent updates to maintain accuracy
Janssens et al. [43]	Unsupervised learning, multi-scale neural network, UMM Discovery	BBBC021 dataset	No need for labeled data, handles batch effects, discovers novel modes of action	Accuracy depends on clustering and normalization methods
Marcelo et al. [45]	ML, DL, RNNs, GANs, VAEs	ChEMBL, DrugBank, PubChem, ZINC, Binding DB	Accelerates antibiotic discovery by predicting antimicrobial activity, de novo molecular design, and identifying drug-likeness.	Relies on the availability and quality of training datasets; computational costs of complex models; limited interpretability of DL methods.
Lin et al. [46]	Repurposing existing drugs, Clinical Trials	COVID-19 clinical data	Repurposing speeds up drug development as safety profiles are already known;	Limited effectiveness against new variants; requires high-quality

ISSN: 2073-9524
Pages:67-85

			lowers the cost and time of development.	preclinical data to show efficacy against SARS-CoV-2.
Kasaraneni [47]	ML, DL, Natural Language Processing (NLP), Predictive Modeling	ChEMBL, DrugBank, PubChem, ZINC, Omics Data	Enhances speed and accuracy of virtual screening, reduces costs, and enables identification of novel drug-disease associations.	Requires high-quality, diverse datasets; limited interpretability of predictions; challenges with data heterogeneity.
Isert et al. [48]	Geometric DL, 3D Graph Neural Networks, SE (3)-Equivariant Neural Networks	ChEMBL Database, PDBBind Dataset	Enabled molecular property prediction, binding pose estimation, and de novo drug design; improved interpretability and prediction accuracy.	High computational requirements; dependency on high-quality 3D molecular datasets; limited generalizability in small datasets.
Krentzel et al. [49]	DL, CNNs, Multiscale CNNs, Transfer Learning	High-content cellular imaging datasets (e.g., BBBC021, Cell Painting datasets, ImageNet)	Enabled accurate mode-of-action (MoA) prediction, robust phenotypic classification, and hit identification from image-based assays.	Requires large, high-quality annotated datasets; computationally expensive; challenges in model interpretability.
Zamir ski et al. [50]	Class-Guided Diffusion Models, Adaptive Group Normalization (AdaGN), Classifier Guidance (CG)	JUMP-CP Cell Painting Dataset, AWS Cell Painting Gallery	Achieved high-quality fluorescent microscopy image generation from brightfield images, improving biological feature quality and interpretability for drug discovery tasks.	Computationally expensive, requiring over 500 GPU hours per plate; high dependency on metadata quality and labeling consistency.
Farrugia et al. [51]	Knowledge Graph Embeddings (ComplEx, LSTM), Graph Autoencoders	DrugBank version 5.1.8	Improved DDI prediction accuracy, handles multiple drug properties, scalable	The complexity of knowledge graph creation, and embedding dimensionality affects performance
Tang et al. [52]	DL, CNNs, Transfer Learning, GANs, VAEs	BBBC, RxRx, JUMP-CP, CPJUMP1, CytoImageNet datasets	Enhanced morphological profiling for drug discovery, improved MOA prediction, and reduced manual feature engineering efforts.	Computationally intensive; limited by dataset quality and size; challenges in integrating multimodal data for comprehensive insights.
Ge et al. [53]	DL (Interaction Transformer Net, ITN), template-based modeling	Protein-peptide interaction datasets (e.g., SH3, MHC I)	High prediction accuracy, combines structural and sequence information, aids peptide drug development	Computationally expensive, reliance on accurate structural data
Ramin et al. [54]	Mechanistic Models, Data-Driven Models, Hybrid Models, Digital Twin	Various bioprocess modeling datasets, Inno4Vac project datasets	Improved scalability, process understanding, and regulatory compliance; reduced development time and costs.	High computational complexity; limited by the availability of high-quality and representative datasets.
Obaido et al. [55]	Supervised Machine Learning (e.g., RF, SVM,	ChEMBL, molecular descriptors, SARS-	Enhanced accuracy in molecular property prediction; accelerated drug	High computational demands; requires high-quality, labeled datasets;

ISSN: 2073-9524
Pages:67-85

	GNNs)	CoV-2 datasets	discovery pipelines; robust predictive capabilities in various applications.	challenges with interpretability in complex models.
Karampuri et al. [56]	Multimodal Deep Neural Networks (MM-DNN), QSAR modeling, Autoencoders, K-means clustering	GDSC (Genomics of Drug Sensitivity in Cancer), Cell Model Passports, PubChem	Achieved high R^2 (0.917) and low RMSE (0.312); facilitated prediction of effective RTK signaling drugs for breast cancer; streamlined molecular profiling integration.	Computationally expensive; dependency on high-quality datasets; potential biases in multimodal data integration.
Hua et al. [57]	Pretraining with CytoImageNet (EfficientNetB0), Transfer Learning, k-NN classification	CytoImageNet (890,737 microscopy images from 40 datasets, including BBBC, IDR, Recursion, and Kaggle datasets)	Improved transfer learning for microscopy image classification; closer domain match for biological tasks compared to ImageNet.	Computationally intensive pretraining; lower generalization performance compared to ImageNet on some tasks.

6. Conclusion

Several future directions in DL for drug discovery focus on leveraging technology to enhance efficiency, accuracy, and personalization. Digital Twinning (DT) enables the simulation of biological systems and more accurate predictions of drug effects, streamlining drug development. Integration of Explainable AI (XAI) addresses the "black-box" nature of DL by creating interpretable models, thereby increasing trust and usability in clinical and regulatory contexts. Advanced Data Integration incorporates multi-omics data, such as genomics, transcriptomics, and proteomics, to build comprehensive models for predicting drug sensitivity, side effects, and efficacy.

Additionally, Novel deep-learning architectures are specifically designed for drug discovery applications. Collaboration across sectors—bringing together pharmaceutical companies, AI firms, and academic institutions—will play a vital role in driving innovation and overcoming technological challenges. Together, these advancements present groundbreaking opportunities to revolutionize drug discovery. Deep learning is revolutionizing drug discovery by introducing advanced methodologies like Graph Neural Networks (GNNs) and Generative Adversarial Networks (GANs). These approaches are energetic key applications, including virtual screening, de novo drug design, protein-ligand interaction prediction, toxicity assessment, and drug repurposing. While challenges such as example data scarcity, noise, and model interpretability remain, strategies like explainable AI (XAI) and multi-task learning are helping to overcome these obstacles. Future advancements, such as integrating multi-omics data and utilizing few-shot learning, hold great promise for further enhancing drug discovery. By enabling more precise predictions, personalized medicine, and efficient molecular exploration, deep learning is transforming pharmaceutical research, optimizing drug development pipelines, and paving the way for groundbreaking new treatments.

Conflict of interest

The authors confirm that there are no conflicts of interest related to the publication of this manuscript.

Reference

[1] Nag, S., Baidya, A. T., Mandal, A., Mathew, A. T., Das, B., Devi, B., & Kumar, R. (2022). Deep learning tools for advancing drug discovery and development. *3 Biotech, 12*(5), 110. https://doi.org/10.1007/s13205-022-03165

[2] 8Waleed, J., Albawi, S., Flayyih, H. Q., & Alkhayyat, A. (2021, September). An effective and accurate CNN model for detecting tomato leaves diseases. In *2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA)* (pp. 33-37). IEEE.

[3] Waleed, J., Azar, A. T., Albawi, S., Al-Azzawi, W. K., Ibraheem, I. K., Alkhayyat, A., ... & Kamal, N. A. (2022). An effective deep learning

ISSN: 2073-9524
Pages:66-84

model to discriminate coronavirus disease from typical pneumonia. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 13*(1), 1-16. https://doi.org/10.4018/IJSSMET.313175

[4] Kolaib, R. J., & Waleed, J. (2024). Crime activity detection in surveillance videos based on a developed deep learning approach. *Diyala Journal of Engineering Sciences, 17*(3), 98–114. https://doi.org/10.24237/djes.2024.17307

[5] Chakraborty, C., Bhattacharya, M., Lee, S.-S., Wen, Z.-H., & Lo, Y.-H. (2024). The changing scenario of drug discovery using AI to deep learning: Recent advancement, success stories, collaborations, and challenges. *Molecular Therapy - Nucleic Acids, 35*(3), 102295. https://doi.org/10.1016/j.omtn.2024.102295

[6] Wu, H., Liu, J., Zhang, R., Lu, Y., Cui, G., Cui, Z., & Ding, Y. (2024). A review of deep learning methods for ligand based drug virtual screening. *Fundamental Research.* https://doi.org/10.1016/j.fmre.2024.02.011

[7] Wu, T., Lin, R., Cui, P., Yong, J., Yu, H., & Li, Z. (2024). Deep learning-based drug screening for the discovery of potential therapeutic agents for Alzheimer's disease. *Journal of Pharmaceutical Analysis, 14*(10), 101022. https://doi.org/10.1016/j.jpha.2024.08.001

[8] Garg, V. (2024). Generative AI for graph-based drug design: Recent advances and the way forward. *Current Opinion in Structural Biology, 84*, 102769. https://doi.org/10.1016/j.sbi.2024.09.003

[9] Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papa Diamantis, A. G., Aidinis, V., ... & Melagraki, G. (2021). Advances in de novo drug design: from conventional to machine learning methods. *International Journal of Molecular Sciences, 22*(4), 1676. https://doi.org/10.3390/ijms22041676

[10] Karki, N., Verma, N., Trozzi, F., Tao, P., Kraka, E., & Zoltowski, B. (2021). SSNet: A deep learning approach for protein-ligand interaction prediction. *International Journal of Molecular Sciences, 22*(3), 1–16. https://doi.org/10.3390/ijms22030116

[11] Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., ... & Pirmohamed, M. (2019). Drug repurposing: Progress, challenges, and recommendations. *Nature Reviews Drug Discovery, 18*(1), 41–58. https://doi.org/10.1038/s41573-018-0004-2

[12] Tanoli, Z., Vähä-Koskela, M., & Aittokallio, T. (2021). Artificial intelligence, machine learning, and drug repurposing in cancer. Expert Opinion on Drug Discovery, 16(9), 977–989. https://doi.org/10.1080/17460441.2021.1912871

[13] Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings, 3*(1), 91–99. https://doi.org/10.1016/j.gtp.2022.02.001

[14] Bhati, M., Virmani, T., Kumar, G., & Sharma, A. (2023). Deep learning in drug discovery. September 2024 Edition.

[15] Lavecchia, A. (2019). Deep learning in drug discovery: Opportunities, challenges, and future prospects. *Drug Discovery Today, 24*(10), 2017–2032. https://doi.org/10.1016/j.drudis.2019.04.021

[16] Schroedl, S. (2020). Current methods and challenges for deep learning in drug discovery. *Drug Discovery Today: Technologies, 32–33*, 9–17. https://doi.org/10.1016/j.ddtec.2020.02.003

[17] Kwon, S., Bae, H., Jo, J., & Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics, 20*(1), 1–12. https://doi.org/10.1186/s12859-019-3039-x

[18] Zhou, Y., Huang, Z., Li, W., Wei, J., Jiang, Q., Yang, W., & Huang, J. (2023). Deep learning in preclinical antibody drug discovery and development. *Methods, 218*, 57–71. https://doi.org/10.1016/j.ymeth.2022.12.007

[19] Lavecchia, A. (2024). Advancing drug discovery with deep attention neural networks. *Drug Discovery Today, 29*(8), 104067. https://doi.org/10.1016/j.drudis.2024.03.001

[20] Bedraoui, A., Suntravat, M., El Mejjad, S., Enezari, S., Oukkache, N., Sanchez, E. E., ... & Daouda, T. (2024). Therapeutic potential of snake venom: Toxin distribution and

ISSN: 2073-9524
Pages:66-84

opportunities in deep learning for novel drug discovery. *Medicine in Drug Discovery, 21*, 100175. https://doi.org/10.1016/j.medidd.2024.100175

[21] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. Springer International Publishing, 25(3).

[22] Kim, J., Park, S., Min, D., & Kim, W. (2021). Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences, 22*(18). https://doi.org/10.3390/ijms22189779

[23] Elbadawi, M., Gaisford, S., & Basit, A. W. (2021). Advanced machine-learning techniques in drug discovery. *Drug Discovery Today, 26*(3), 769–777. https://doi.org/10.1016/j.drudis.2020.11.014

[24] van Tilborg, D., Brinkmann, H., Criscuolo, E., Rossen, L., Özçelik, R., & Grisoni, F. (2024). Deep learning for low-data drug discovery: Hurdles and opportunities. *Current Opinion in Structural Biology, 86*, 102818. https://doi.org/10.1016/j.sbi.2024.02.004

[25] Lin, M., Cai, J., Wei, Y., Peng, X., Luo, Q., Li, B., ... & Wang, L. (2024). MalariaFlow: A comprehensive deep learning platform for multistage phenotypic antimalarial drug discovery. *European Journal of Medicinal Chemistry, 277*, 116776. https://doi.org/10.1016/j.ejmech.2024.116776

[26] Islam, T., Hafiz, M. S., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthcare Analytics, 5*, 100340. https://doi.org/10.1016/j.hcan.2024.100340

[27] Tarasova, O. A., Rudik, A. V., Ivanov, S. M., Lagunin, A. A., Poroikov, V. V., & Filimonov, D. A. (2021). Machine learning methods in antiviral drug discovery. *Topics in Medicinal Chemistry, 37*, 245–279. https://doi.org/10.1007/s11306-021-00667-3

[28] Wu, J., Xiao, Y., Lin, M., Cai, H., Zhao, D., Li, Y., ... & Wang, L. (2023). DeepCancerMap: A versatile deep learning platform for target-and cell-based anticancer drug discovery. *European Journal of Medicinal Chemistry, 255*, 115401. https://doi.org/10.1016/j.ejmech.2023.115401

[29] Shimizu, Y., Yonezawa, T., Bao, Y., Sakamoto, J., Yokogawa, M., Furuya, T., ... & Ikeda, K. (2023). Applying deep learning to iterative screening of medium-sized molecules for protein–protein interaction-targeted drug discovery. *Chemical Communications, 59*(44), 6722-6725. https://doi.org/10.1039/D3CC01647E

[30] Liang, P. P., Wu, P., Ziyin, L., Morency, L. P., & Salakhutdinov, R. (2021). Cross-modal generalization: Learning in low-resource modalities via meta-alignment. *Proceedings of the 29th ACM International Conference on Multimedia (MM 2021)*, 2680–2689. https://doi.org/10.1145/3474085.3475300

[31] Chen, X., Lin, Y., Liu, M., & Gilson, M. K. (2002). The binding database: Data management and interface design. *Bioinformatics, 18*(1), 130–139. https://doi.org/10.1093/bioinformatics/18.1.130

[32] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research, 35*(SUPPL. 1), 198–201. https://doi.org/10.1093/nar/gkl200

[33] Wang, Y., Cheng, T., & Bryant, S. H. (2017). PubChem BioAssay: A decade's development toward open high-throughput screening data sharing. *SLAS Discovery, 22*(6), 655–666. https://doi.org/10.1177/2472555217710880

[34] Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W., Kowalczyk, W., ... & Van Westen, G. J. (2017). Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics, 9*, 1-14. https://doi.org/10.1186/s13321-017-0262-6

[35] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery*

ISSN: 2073-9524
Pages:66-84

Today, *23*(6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039

[36] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., ... & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology, 37*(9), 1038–1040. https://doi.org/10.1038/s41587-019-0238-6

[37] Shtar, G., Rokach, L., & Shapira, B. (2019). Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. *PLoS One, 14*(8), 1–21. https://doi.org/10.1371/journal.pone.0221052

[38] Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... & Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research,* *47*(D1), D930–D940. https://doi.org/10.1093/nar/gky1072

[39] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., ... & Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research, 47*(D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033

[40] Jacquemard, C., & Rognan, D. (2022). LIT-PCBA: An unbiased dataset for machine learning and virtual screening. HAL Id: hal-03013622.

[41] Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., ... & Leach, A. R. (2020). An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics,* *12*, 1-16. https://doi.org/10.1186/s13321-020-00456-2

[42] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., ... & Bolton, E. E. (2021). PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research,* *49*(D1), D1388-D1395. https://doi.org/10.1093/nar/gkaa1055

[43] Janssens, R., Zhang, X., Kauffmann, A., De Weck, A., & Durand, E. Y. (2021). Fully unsupervised deep mode of action learning for phenotyping high-content cellular images. *Bioinformatics,* *37*(23), 4548–4555. https://doi.org/10.1093/bioinformatics/btab623

[44] Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y. A. M. M., Gomaa, M. M., & Hassanien, A. E. (2023). Deep learning in drug discovery: An integrative review and future challenges. Springer Netherlands, 56(7).

[45] Melo, M. C. R., Maasch, J. R. M. A., & de la Fuente-Nunez, C. (2021). Accelerating antibiotic discovery through artificial intelligence. *Communications Biology, 4*(1), 1–13. https://doi.org/10.1038/s41522-021-00341-z

[46] Lin, M., Dong, H. Y., Xie, H. Z., Li, Y. M., & Jia, L. (2021). Why do we lack a specific magic anti-COVID-19 drug? Analyses and solutions. *Drug Discovery Today,* *26*(3), 631–636. https://doi.org/10.1016/j.drudis.2020.12.012

[47] Kasaraneni, R. K. (2021). AI-enhanced virtual screening for drug repurposing: Accelerating the identification of new uses for existing drugs, 1(2), 129–161.

[48] Isert, C., Atz, K., & Schneider, G. (2023). Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology,* *79*, 1–10. https://doi.org/10.1016/j.sbi.2023.07.006

[49] Krentzel, D., Shorte, S. L., & Zimmer, C. (2023). Deep learning in image-based phenotypic drug discovery. *Trends in Cell Biology,* *33*(7), 538–554. https://doi.org/10.1016/j.tcb.2023.04.002

[50] Cross-Zamirski, J. O., Anand, P., Williams, G., Mouchet, E., Wang, Y., & Schönlieb, C. B. (2023). Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. *Proceedings of the 2023 IE*EE/CVF International Conference on Computer Vision Workshops (ICCVW 2023), 3802–3811. https://doi.org/10.1109/ICCVW56367.2023.00384

[51] Farrugia, L., Azzopardi, L. M., Debattista, J., & Abela, C. (2023). Predicting drug-drug interactions using knowledge graphs.

[52] Wu, H., Liu, J., Zhang, R., Lu, Y., Cui, G., Cui, Z., & Ding, Y. (2024). A review of deep learning methods for ligand based drug virtual screening. *Fundamental Research*.

ISSN: 2073-9524
Pages:66-84

[53] Ge, J., Jiang, D., Sun, H., Kang, Y., Pan, P., Deng, Y., ... & Hou, T. (2024). Deep-learning-based prediction framework for protein-peptide interactions with structure generation pipeline. *Cell Reports Physical Science, 5*(6). https://doi.org/10.1016/j.xcrp.2024.100541

[54] Ramin, E., Cardillo, A. G., Liebers, R., Schmölder, J., Von Lieres, E., Van Molle, W., ... & Gernaey, K. V. (2024). Accelerating vaccine manufacturing development through model-based approaches: Current advances and future opportunities. *Current Opinion in Chemical Engineering, 43*, 100998. https://doi.org/10.1016/j.coche.2024.100998

[55] Obaido, G., Mienye, I. D., Egbelowo, O. F., Emmanuel, I. D., Ogunleye, A., Ogbuokiri, B., ... & Aruleba, K. (2024). Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Machine Learning with Applications, 17*, 100576. https://doi.org/10.1016/j.mlwa.2024.100576

[56] Karampuri, A., Kundur, S., & Perugu, S. (2024). Exploratory drug discovery in breast cancer patients: A multimodal deep learning approach to identify novel drug candidates targeting RTK signaling. *Computers in Biology and Medicine, 174*(February), 108433. https://doi.org/10.1016/j.compbiomed.2024.108433

[57] Hua, S. B. Z., Lu, A. X., & Moses, A. M. (2021). CytoImageNet: A large-scale pretraining dataset for bioimage transfer learning. NeurIPS Conference Proceedings.

[58] Blanco-Gonzalez, A., Cabezon, A., Seco-Gonzalez, A., Conde-Torres, D., Antelo-Riveiro, P., Pineiro, A., & Garcia-Fandino, R. (2023). The role of AI in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals, 16*(6), 891. https://doi.org/10.3390/ph16060891