Current Proteomics 22 (2025) 100015

Contents lists available at ScienceDirect

Current Proteomics

journal homepage: http://www.keaipublishing.com/en/journals/current-proteomics/

Advances in machine learning for optimizing pharmaceutical drug discovery

Uzochukwu Eric Ogbonna [a], Ezichi Favour Ofoezie [b], Olorunfemi Oyewole Babalola [c,*], Paul Olamide Ottu [c], Chinwendu Angela Ogbonna [d], Sandra Olisakwe [e], Tiffany Ezinne George [f], Simeon Babarinde [g], Justina Onyinyechi Omaba [h], Chidera Godson Chukwuemeka [i], Chibuzo Collette Amafili [j], Anioke Afamefuna [k], Henry Ogbonna [a]

[a] Department of Biochemistry, College of Natural Sciences, Michael Okpara University of Agriculture, Umudike, Umuahia, Nigeria
[b] Department of Microbiology, Federal Polytechnic, Nekede, Owerri, Nigeria
[c] Teady Bioscience Research Laboratory, Ilaramokin, Ondo State, Nigeria
[d] Division of Cardiovascular Disease, University of Alabama at Birmingham, Birmingham, USA
[e] Department of Anatomy, University of Nigeria Nsukka, Enugu, Nigeria
[f] Department of Hematology and Oncology, University of Alabama at Birmingham, Birmingham, USA
[g] Department of Biology and Biochemistry, University of Bath. Bath, UK
[h] Department of Medical Laboratory Science, Nnamdi Azikiwe University, Nnewi, Nigeria
[i] Department of Mathematics, Imo State University, Owerri. Nigeria
[j] Department of Nursing, Ebonyi State University, Abakaliki, Nigeria
[k] Department of Microbiology, Ebonyi State University, Abakaliki, Nigeria

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Transfer learning
Federated learning
Drug discovery

ABSTRACT

The integration of advanced machine learning methodologies has revolutionized the field of pharmaceutical drug discovery by addressing critical challenges in efficiency, scalability, and accuracy. This manuscript explores key machine learning paradigms, including deep learning, natural language processing, transfer learning, few-shot learning, and federated learning, highlighting their transformative impact on drug discovery processes. Deep learning architectures, such as convolutional neural networks, recurrent neural networks, and attention-based models, have enabled precise predictions of molecular properties, protein structures, and ligand-target interactions, accelerating lead compound identification and optimization. Natural language processing tools like SciBERT and BioBERT have streamlined biomedical knowledge extraction, uncovering novel drug-disease relationships and facilitating rapid therapeutic advancements during critical scenarios such as the COVID-19 pandemic. Transfer learning and few-shot learning have proven effective in scenarios with limited datasets, leveraging pre-trained models to predict molecular properties, optimize lead compounds, and identify toxicity profiles. Meanwhile, federated learning has enabled secure multi-institutional collaborations, integrating diverse datasets to discover biomarkers, predict drug synergies, and enhance virtual screening without compromising data privacy. These methodologies collectively foster innovation in drug discovery by enhancing precision, reducing development timelines, and lowering costs. By providing an overview of these cutting-edge machine learning approaches and their applications, this review underscores their potential to accelerate pharmaceutical research, optimize therapeutic development, and address complex challenges in the evolving landscape of drug discovery.

* Corresponding author.
E-mail address: Babalolaolorunfemi20@gmail.com (O.O. Babalola).

https://doi.org/10.1016/j.curpro.2025.100015
Received 31 January 2025; Received in revised form 25 April 2025; Accepted 6 July 2025
Available online 26 July 2025
1570-1646/© 2025 The Authors. Publishing Services by Elsevier B.V. on behalf of KeAi. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

The drug discovery process is fundamental in advancing human health but is fraught with significant challenges. Traditional drug discovery involves a multi-step process, including target identification, lead discovery, optimization, preclinical testing, and clinical trials. This process is not only lengthy, often requiring over a decade, but also financially burdensome, with the average cost of bringing a new drug to market exceeding $2 billion. The failure rate is also alarmingly high, with only about 10 % of candidates that enter clinical trials ultimately achieving market approval.[1,2] These inefficiencies are driven by several factors, including the complexity of biological systems, challenges in predicting human pharmacokinetics and pharmacodynamics, and limitations in preclinical models. As a result, there is an urgent need for innovative strategies to improve the speed, efficiency, and accuracy of drug discovery efforts.

Over the past decade, machine learning (ML), a subfield of artificial intelligence (AI), has gained prominence as a transformative tool in pharmaceutical research. ML leverages computational algorithms capable of learning patterns from vast datasets, enabling predictions, decision-making, and the discovery of novel insights without explicit programming.[3] The integration of ML into drug discovery has been catalyzed by advancements in computational power, availability of large-scale biological and chemical datasets, and innovations in algorithmic design, such as deep learning.[4] Unlike traditional computational methods, ML excels in handling the complexity and heterogeneity of biomedical data, which often spans genomics, proteomics, cheminformatics, and clinical records. The applications of ML in drug discovery are manifold. For example, ML algorithms have been employed to enhance target identification by analyzing omics data to predict novel therapeutic targets with high confidence.[5] In lead discovery, ML-based virtual screening methods have revolutionized the identification of small molecules with desired biological activity, significantly reducing the reliance on experimental high-throughput screening[6]. Furthermore, ML has been applied to de novo drug design, where generative models create entirely new chemical structures with optimized pharmacological properties.[7] In the preclinical phase, ML is transforming in silico toxicology by predicting adverse effects based on molecular structure and biological activity. Models trained on comprehensive datasets, such as Tox21 and PubChem, have been used to assess cardiotoxicity, hepatotoxicity, and other safety concerns with high reliability.[8,9] In clinical trials, ML has been applied to optimize trial design, predict patient enrollment rates, and identify biomarkers for patient stratification. For instance, clustering algorithms have been employed to segment patient populations based on genetic or phenotypic data, facilitating precision medicine approaches. Additionally, reinforcement learning models have been explored to simulate trial scenarios and identify strategies that maximize the likelihood of success.[10] Recent advances in ML have focused on integrating multimodal data from diverse sources, such as omics, imaging, and electronic health records, to create comprehensive models of disease and drug response. The development of interpretable ML methods, such as attention mechanisms and explainable AI frameworks, has addressed concerns regarding model transparency, fostering greater adoption in pharmaceutical R&D. Moreover, collaborative initiatives, such as the Open Targets platform, are leveraging ML to accelerate target validation and drug repurposing efforts.[11] As such, this review is aimed at the exploration of ML applications across the drug discovery continuum, highlighting their transformative potential in pharmaceutical innovation.

2. Machine learning applications in key stages of drug discovery

2.1. Target identification and validation

The identification of gene-disease relationships is a foundational step in drug discovery, enabling the prediction of therapeutic targets with high confidence. ML methods excel in analyzing vast genomic datasets, identifying patterns, and establishing causal relationships between genes and diseases. For example, decision tree (DT) classifiers, as demonstrated by Costa et al,[12] utilize features such as mRNA expression, gene essentiality, mutation frequency, and protein-protein interactions to predict morbidity-associated and druggable genes. Their model achieved a notable recovery rate of 65 % for known morbid genes and 78 % for unknown druggable genes, thereby reducing the reliance on labor-intensive experimental procedures and accelerating the identification of key molecular targets (Table 1). Furthermore, ML techniques like support vector machines (SVMs) have been deployed to analyze genomic variability and classify proteins based on their potential for drug binding. Jeon et al[13] successfully used this approach in breast, pancreatic, and ovarian cancers, identifying 122 global cancer targets, with some predictions validated through peptide inhibitors exhibiting antiproliferative effects. These findings underscore the role of ML in extracting actionable insights from complex genomic data to prioritize potential targets for therapeutic development.

Biomarkers play a critical role in diagnosing diseases, predicting therapeutic responses, and monitoring treatment efficacy. Machine learning models have proven instrumental in discovering novel biomarkers by processing high-dimensional omics data, including transcriptomics and proteomics. Mamoshina et al[14] applied ML to identify biomarkers associated with human aging, analyzing gene expression data from the GTEx project. Their SVM-based approach achieved an accuracy level of 0.80, revealing significant biomarkers linked to muscle tissue aging. By leveraging supervised learning techniques and deep feature selection, ML enables precise identification of biomarkers, facilitating the development of targeted therapeutic interventions. Predictive biomarker discovery, in particular, has become a cornerstone of precision medicine. ML-based predictive models not only enhance clinical trial success rates but also shed light on drug mechanisms of action and facilitate patient stratification.[15–17] Late-stage clinical trials demand immense resources and time; hence, applying ML models to preclinical or early-stage clinical data offers a more efficient pathway. Translational biomarkers can be predicted using ML techniques on preclinical datasets, and once validated with independent datasets, these models can stratify patients, propose potential drug indications, and identify mechanisms of action.[15] However, the integration of these biomarkers into clinical trials has faced challenges, such as issues related to data quality, model reproducibility, and the design of clinically compatible assays. Community-driven initiatives have aimed to address these gaps. The MicroArray Quality

Table 1
Applications and advances in AI-driven drug repurposing: techniques and case studies.

Stages of Drug Discovery	Subsection	Description	Case Studies/Examples	References
Target Identification and Validation	Gene-Disease Associations	Machine learning (ML) helps establish causal relationships between genes and diseases by analyzing genomic datasets, predicting therapeutic targets with high confidence, and reducing reliance on experimental methods.	- Decision tree classifiers achieved a 65% recovery rate for known morbid genes and 78% for unknown druggable genes. - SVM models identified 122 global cancer targets.	12,13
	ML for Biomarker Discovery	ML models process high-dimensional omics data to discover novel biomarkers for disease diagnosis, treatment response prediction, and monitoring treatment efficacy.	- SVM-based approach identified biomarkers linked to muscle tissue aging with 80% accuracy. - MAQC II project evaluated various ML methods for predicting clinical endpoints.	14,18
	Predictive Biomarker Discovery	ML-based predictive models enhance clinical trial success rates, patient stratification, and drug mechanism identification, providing insights into translational biomarkers.	- ML techniques stratified patients and proposed drug indications using preclinical datasets. - Cox regression models predicted high-risk subgroups in multiple myeloma.	15,19,22-23
	Target Validation	Validation remains a bottleneck in drug discovery. ML helps predict clinical trial success by integrating genetic and expression data to prioritize disease-specific targets for validation.	- Deep learning models predicted alternative splicing signals and their disease associations. - Rouillard et al. predicted clinical trial outcomes using omics data.	27,29, 30
	Integration of Multi-Omics Data	Modern biology generates multidimensional datasets. ML models integrate genomics, proteomics, and imaging to uncover relationships that traditional methods overlook, improving prediction robustness.	- Ferrero et al. identified genetic data and animal models as critical predictors for therapeutic target success. - NLP tools (BeFree, DigSee) mine literature databases for drug-target data.	26
	Application in Cancer Target Discovery	ML has identified disease-specific targets across cancers such as breast, pancreatic, and ovarian, validating predictions experimentally through inhibition of cancer cell viability.	- SVM models classified proteins as drug targets in breast, pancreatic, and ovarian cancer. - ML predicted drug responses using gene expression and mutation data from 990 cancer cell lines.	13, 24
	Application in Neurodegenerative Diseases	Regression models identified transcriptional changes in Huntington's disease, reconstructing a list of target genes for transcription factors in the mouse striatum.	- LASSO regularization revealed differential expression patterns in disease tissues, which could guide therapeutic interventions in Huntington's disease.	25
Hit Discovery and Compound Screening	Virtual Screening (VS)	Virtual screening is a computational approach to explore large compound libraries and identify potential drug candidates, reducing the cost and time of traditional screening. It allows prioritization of high-potential molecules for further testing.	Valentini et al. (2014) highlighted the role of VS in expediting hit identification, while Ferrero et al. (2017) explored its integration into modern drug discovery pipelines.	26, 31
	High-Throughput Virtual Screening (HTVS)	HTVS utilizes computational methods to rapidly identify compounds that interact with specific drug targets. AI and machine learning (ML) models, such as Support Vector Machines (SVMs) and Random Forests (RFs), enhance prediction accuracy.	Saeh et al. (2005) achieved a 69x higher success rate using SVM for GPCR activity screening. Franke et al. (2005) identified 11 hit compounds from 225,000 molecules targeting dopamine receptors.	34-35
	Support Vector Machines (SVMs)	SVMs classify compounds using kernel functions to map nonlinear molecular descriptors, enabling accurate analysis of chemical interactions. This approach has improved hit discovery across various therapeutic areas.	Burbidge et al. (2001) applied SVMs to the NCI AIDS dataset, outperforming traditional classification methods. Saeh et al. (2005) screened 130,000 compounds for GPCR activity with high success.	32-34
	ML-Based Scoring Functions	Scoring functions in molecular docking predict binding affinity. ML-based functions such as ΔvinaRF improve predictive performance by integrating experimental binding data with traditional docking scores.	Wang and Zhang (2017) introduced ΔvinaRF, combining RF and AutoDock scoring. Jimenez et al. (2018) developed a CNN-based model for predicting protein-ligand binding affinities with high precision.	36, 100

(*continued on next page*)

Table 1 (*continued*)

Stages of Drug Discovery	Subsection	Description	Case Studies/Examples	References
	Deep Generative Models (GANs and VAEs)	Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are used to create novel molecular structures tailored to specific pharmacological profiles, improving chemical diversity and de novo molecular design.	Kadurin et al. (2017) demonstrated the use of GANs in designing anticancer agents. Gómez-Bombarelli et al. (2018) introduced the "Molecular Autoencoder" for molecular optimization and scaffold design.	40, 43
	Integration of GANs and VAEs with ML	Combining GANs and VAEs with ML techniques such as recurrent neural networks (RNNs) enables sequential molecular design, similar to natural language processing methods.	Perron et al. (2022) applied this approach to discover inhibitors for CDK4 and PIM1 kinases, confirming the activity of the lead compounds through experimental validation.	45-46
	Active Learning	Active learning iteratively refines ML models by selecting and prioritizing the most informative compounds for evaluation, optimizing resource allocation and improving the efficiency of virtual screening workflows.	Franke et al. (2005) identified 11 potent hits using active learning for dopamine receptor activity screening. Wang and Zhang (2017) improved molecular docking accuracy through active learning.	35-36
	Bayesian Optimization in Active Learning	Bayesian optimization focuses on uncertain predictions, refining models iteratively to improve accuracy and efficiency in hit discovery processes.	Assmann et al. (2019) discussed the challenges and potential opportunities in integrating Bayesian optimization with active learning for compound screening workflows.	47
Lead Optimization	Lead Optimization	Optimization of lead compounds requires assessing ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties. ML and AI methods such as RF, SVM, ANN, and DL are used to predict these properties to enhance drug discovery.	- Ye et al. (2018) used transfer and multitask learning to predict human pharmacokinetics with an MAE of 0.31 for oral bioavailability. - Lou et al. (2021) developed models predicting subcutaneous bioavailability for monoclonal antibodies with 78% accuracy. - Obrezanova et al. (2023) applied convolutional neural networks for computing in vivo pharmacokinetics of 3000+ compounds. - DeepTox predicted toxicological risks in novel compounds.	48-52, 56-61
	QSAR Modeling in Drug Discovery	QSAR modeling predicts biological activity and toxicity of compounds based on chemical structures. It aids lead optimization by correlating chemical properties with biological activity. Modern ML/DL algorithms improve QSAR accuracy and streamline lead selection while reducing experimental costs.	- Dahl et al. (2014) showed multi-task DNNs outperform single-task networks in feature recognition. - Ma et al. (2015) used ensemble models combining DNN and gradient boosting for predicting binding affinities. - Williams et al. (2019) applied ML to predict drug-induced liver injury (DILI) with high sensitivity and specificity. - Verma and Matthews (2015) achieved >95% accuracy in predicting eye and skin irritation with QSTR-ANN models. - Mamoshina et al. (2020) integrated transcriptional profiling with QSAR for cardiotoxicity prediction.	62-64, 66, 68-69, 72-74
	Molecular Docking & ML Integration	Molecular docking, combined with ML, enhances virtual screening and lead optimization by improving binding affinity predictions and refining scoring functions. ML-driven docking improves accuracy and computational efficiency.	- Gupta et al. (2020) used ML for identifying HDAC inhibitors in AD. - Chen et al. (2019) applied DL to predict bioactivity of GSK3β inhibitors. - Jamal et al. (2019) predicted caspase-8 inhibitors using molecular dynamics-based descriptors. - Shao et al. (2018) developed dual antagonists for A2A adenosine and dopamine D2 receptors using SVM and in vitro validation. - Deng et al. (2020) used ML to predict bioactivity of Galectin-3 inhibitors in Huntington's disease. - Zeng et al. (2019) developed deepDR for drug repurposing in AD and PD.	75-79, 82-84

(*continued on next page*)

Table 1 (*continued*)

Stages of Drug Discovery	Subsection	Description	Case Studies/Examples	References
Drug Repurposing	Drug Repurposing	Drug repurposing involves identifying new therapeutic applications for existing drugs, significantly reducing time and cost compared to traditional drug development. This strategy bypasses early development stages and progresses directly to advanced clinical testing. Machine learning (ML) and artificial intelligence (AI) have revolutionized drug repurposing by analyzing large datasets and discovering novel drug-disease associations.	Risperidone and aripiprazole identified for Alzheimer's disease; methylphenidate and pergolide for Parkinson's disease using deep learning (Zeng et al., 2019). DTINet predicted novel cyclooxygenase inhibitors like alendronate and chlorpropamide for anti-inflammatory effects (Luo et al., 2017). During COVID-19, AI models identified potential anti-HCoV drugs and stage II breast cancer therapies (Hooshmand et al., 2020; Fahimian et al., 2019).	83, 85-86, 92-93, 95
	Machine Learning Techniques	ML techniques like Support Vector Machines (SVM), Neural Networks (NN), and Deep Learning (DL) are pivotal in drug repurposing. SVM classifies drug-target interactions, particularly in network-based approaches. Neural Networks capture complex, non-linear relationships, integrating data from genomics, proteomics, and chemical structures. Deep Learning further enhances drug repurposing by uncovering intricate patterns in biological networks and predicting drug candidates.	SVM models predict therapeutic targets using multi-dimensional datasets, including genomic, chemical, and protein interaction data (Koromina et al., 2019). DL models like DeepDR integrate drug-disease networks to predict candidates for repositioning. Examples include fibrosis and cancer therapies using DTI-RCNN, which combines recurrent neural networks with drug-target interaction data (Wu et al., 2020).	87, 90
	Specialized Algorithms	Algorithms like PREDICT, SPACE, and DTINet integrate diverse data sources such as chemical properties, gene expression profiles, and biological networks to identify drug repurposing candidates. Network-based methods model complex relationships among drugs, targets, and diseases.	DTINet integrates multiple biological networks using dimensionality reduction and network diffusion, predicting novel inhibitors for cyclooxygenase. BiFusion enhances prediction accuracy for viral and non-viral diseases (Wang et al., 2020).	91-92, 101
	Real-World Data (RWD)	The incorporation of real-world data (RWD) from electronic health records (EHR), patient registries, and wearable devices provides insights into drug safety, efficacy, and real-world applicability. RWD captures population variability, addressing limitations of homogenous clinical trial cohorts and offering opportunities for drug repurposing in complex diseases.	Drug-disease networks constructed from EHR and genomic data help identify shared molecular mechanisms. During COVID-19, RWD identified potential antiviral candidates. RWD also helped uncover repurposing opportunities for chronic conditions like fibrosis and cancer (Fahimian et al., 2019; Wu et al., 2020).	[90, 95, 98

Control II (MAQC II) project, organized by the FDA, evaluated various ML methods for predicting clinical endpoints using baseline gene expression data.[18] In this study, 36 independent teams analyzed six microarray datasets to develop predictive models for 13 clinical endpoints. Key findings emphasized the critical role of data quality, skilled scientists, and appropriate model selection in improving prediction accuracy. For example, a univariate Cox regression approach in multiple myeloma identified a gene expression signature that effectively predicted a high-risk subgroup of patients without relying on arbitrary survival cut-offs, demonstrating the advantages of regression-based models for continuous variables.[19–23]

ML-driven strategies for target identification have been successfully applied across various diseases and therapeutic areas. For instance, Jeon et al[13] applied an SVM model to classify proteins as drug targets or non-drug targets in cancers. Their predictions identified 266 targets specific to breast cancer, 462 to pancreatic cancer, and 355 to ovarian cancer, with experimental validation showing strong inhibition of cell viability for several predicted targets. This highlights the practical utility of ML in pinpointing disease-specific targets and validating them experimentally. Similarly, Iorio et al[24] integrated ML algorithms with gene expression, DNA methylation, and mutation data to predict drug responses across 990 cancer cell lines. Their approach demonstrated that molecular features such as gene expression, driver mutations, and DNA methylation are crucial predictors of drug efficacy. This methodology has advanced our understanding of cancer-specific drug effects, offering new opportunities for tumor-intrinsic precision medicine. Moreover, Ament et al[25] utilized regression models and LASSO regularization to identify transcriptional changes in Huntington's disease. Their genome-scale model reconstructed a list of target genes for 718 transcription factors in the mouse striatum, revealing differential expression patterns in disease tissues, which could serve as starting points for potential therapeutic interventions.

Modern biology generates multidimensional datasets spanning genomics, proteomics, and imaging. Machine learning models integrate these datasets to uncover intricate relationships that traditional approaches might overlook. Ferrero et al[26] applied ML classifiers using target-disease associations from the Open Targets platform, identifying genetic data and animal models as critical predictors for therapeutic target success. This integration of diverse omics data allows for more robust predictions, although challenges remain in ensuring the availability of comprehensive datasets detailing both successful and failed drug discovery programs. Moreover, the literature remains a vital source of knowledge on target-disease association. Advances in natural language processing, an ML approach for text mining, have enabled automated extraction of relevant data from unstructured text.

While ML has proven effective in identifying and prioritizing targets, the validation of these targets remains a critical bottleneck. Rouillard et al[27] demonstrated the potential of ML to predict clinical trial success based on omics data, emphasizing the importance of integrating genetic and expression data to identify disease-specific targets. As such, the use of deep learning models to predict alternate splicing signals and their association with diseases, as reported by several studies,[28–30] represents a significant milestone. These approaches refined our understanding of complex biological processes and guided the development of more effective and precise therapeutic interventions.

2.2. Hit discovery and compound screening

Virtual screening (VS) is a transformative tool in drug discovery, significantly expediting the identification of potential drug candidates by efficiently exploring large compound libraries. VS has evolved as a critical step in the drug discovery pipeline, enabling researchers to prioritize molecules with high therapeutic potential before committing to costly and time-intensive experimental procedures.[26,31] The integration of artificial intelligence and machine learning approaches has further enhanced the precision and efficiency of VS methodologies. High-Throughput Virtual Screening leverages computational methods to identify compounds that interact with specific drug targets. AI and ML algorithms, such as support vector machines (SVMs) and random forests (RFs), are employed to process vast datasets, predict molecular properties, and rank compounds based on their likelihood of success.

The SVM approach, for instance, has been effectively combined with cheminformatics techniques to classify compounds with high accuracy. Burbidge et al[32] demonstrated that SVM classification using kernel functions outperformed traditional methods in analyzing the National Cancer Institute (NCI) AIDS dataset, identifying compounds with enhanced specificity and activity profiles. The kernel function, a distinguishing feature of SVM, enables nonlinear mapping of molecular descriptors, allowing the analysis of complex chemical and biological interactions.[33] Applications of SVM in HTVS include ranking molecules for G protein-coupled receptor (GPCR) activity, where Saeh et al[34] reported a success rate 69 times higher than random selection by screening 130,000 compounds. Similarly, Franke et al[35] utilized SVM-based virtual screening to identify 11 hit compounds from 225,000 molecules targeting dopamine receptors, achieving strong binding selectivity (Table 1).

The scoring function is a pivotal component in molecular docking, used to predict the binding affinity of hit compounds with their target proteins. Traditional docking scoring functions rely on physical and chemical principles, but ML-based approaches have revolutionized this process by learning complex nonlinear relationships from experimental data. Wang and Zhang[36] introduced the ΔvinaRF scoring function, which combines RF and AutoDock scoring, demonstrating superior predictive performance. ML tools such as RFs and SVMs improve screening power by incorporating diverse features, including geometric, chemical, and physical descriptors. For example, ΔvinaRF integrates experimental binding affinity data with docking scores to enhance lead identification. These methods offer robust performance even when noisy or incomplete data are present, providing a reliable framework for hit validation and optimization.[37]

Several studies have highlighted the potential of ML-based VS in diverse therapeutic areas. Lepp et al[38] screened over 30,000 ligands per dataset related to depression using SVMs and atom-count descriptors, achieving recall rates between 44 % and 89 %. Franke et al[35] demonstrated the effectiveness of ML approaches in screening 2.7 million compounds for Cox-2 antagonists, identifying three potent molecules, including one with superior inhibitory profiles compared to celecoxib and rofecoxib. Similarly, Jorissen and Gilson[39] utilized SVMs with 2D descriptors generated through DRAGON software to screen 25,175 compounds from the Maybridge

database and reported improved performance over traditional methods. These applications underline the versatility and scalability of ML algorithms in tackling large datasets and uncovering novel drug candidates.

The advent of deep generative models, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), has significantly advanced the field of drug discovery by facilitating the creation of novel molecular structures tailored to specific pharmacological profiles. These models address longstanding challenges in chemical diversity, optimization, and de novo molecular design. GANs utilize a dual-network architecture composed of a generator and a discriminator. The generator creates molecular structures, while the discriminator evaluates their authenticity relative to real molecular data, resulting in a dynamic feedback loop that refines the quality of generated molecules. For instance, Kadurin et al. demonstrated the application of GANs for designing selective anticancer agents, emphasizing their capacity to optimize molecular profiles through adversarial training.[40] GANs have also been employed to generate molecules with desired properties such as drug-likeness, binding affinity, and solubility, significantly enhancing early-stage drug discovery processes.[41]

DruGAN, an advanced GAN implementation, exemplifies the potential of these models by enabling the generation of molecules with optimized pharmacokinetic properties. Furthermore, these models are instrumental in creating diverse chemical libraries that align with predefined target attributes, thereby accelerating the identification of lead candidates. The integration of GANs into virtual screening workflows has provided valuable insights into ligand-receptor interactions and enabled the discovery of novel binding sites, particularly for challenging drug targets like GPCRs and allosteric sites.[42] On the other hand, VAEs adopt a distinct approach by encoding molecular data into a latent space representation and decoding it to reconstruct the original structures. This framework facilitates the exploration of chemical space and the generation of novel compounds through manipulation of latent variables.

Gómez-Bombarelli et al. introduced the "Molecular Autoencoder," which integrates an encoder network for transforming chemical structures into continuous latent space vectors and a decoder for reconstructing these into chemical entities.[43] VAEs are particularly advantageous for designing molecules with specific features, such as increased solubility, enhanced bioavailability, or target-specific binding. Recent applications of VAEs have demonstrated their utility in optimizing molecular scaffolds and generating large-scale compound libraries. These libraries have significantly improved hit discovery rates by identifying candidate molecules with favorable pharmacokinetic and pharmacodynamic profiles.[44] Additionally, VAEs provide a robust platform for structure-activity relationship studies, enabling the rational design of drug candidates with enhanced therapeutic efficacy.

The fusion of GANs and VAEs with other machine learning techniques has further elevated their impact in molecular design. For example, combining these generative models with recurrent neural networks has shown promise in de novo molecular design by leveraging sequential information akin to natural language processing methods.[45] This synergy has proven effective in identifying inhibitors for well-studied targets like CDK4 and PIM1 kinases, with experimental validations confirming the activity of lead compounds.[46] Despite their potential, practical challenges remain in applying generative models to drug discovery. Assmann et al. highlighted difficulties in optimizing multiparameter objectives and integrating these models into workflows for high-throughput compound screening.[47] Addressing these challenges through continued development of generative frameworks and integration with cheminformatics tools will further enhance their utility in modern drug discovery.

Active learning has emerged as a transformative approach to compound screening, addressing the cost and time constraints associated with traditional high-throughput screening methods. By focusing on iterative model refinement, active learning prioritizes the evaluation of compounds predicted to yield the most significant information gain, thereby optimizing resource allocation. The principle of active learning lies in its ability to selectively query the most informative data points from a vast chemical space. This iterative process involves training an initial model, evaluating its predictions, and refining it based on experimental feedback from selected compounds.

Machine learning models such as SVMs and RFs have been extensively utilized within active learning frameworks for virtual screening. These models excel in handling high-dimensional data and nonlinear relationships, making them ideal for predicting molecular properties and binding affinities. Franke et al[35] reported the successful application of active learning in screening over 225,000 compounds for dopamine receptor activity, yielding 11 potent hits with high binding selectivity. Similarly, Wang and Zhang[36] incorporated active learning with RF-based scoring functions to enhance the accuracy of molecular docking simulations, resulting in improved identification of lead compounds.

Furthermore, active learning methods have been enhanced by incorporating domain-specific knowledge, such as pharmacophore models and SAR data, to guide the selection process. The integration of active learning with generative models has also shown promise, enabling the design and prioritization of novel compounds that align with therapeutic objectives. Recent advancements in active learning algorithms, such as uncertainty sampling and Bayesian optimization, have further improved the efficiency and accuracy of compound screening workflows.

2.3. Lead optimization

The optimization of lead compounds during drug discovery is a critical stage that demands meticulous evaluation of Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties. Effective absorption is essential for the therapeutic efficacy of drug candidates. ML algorithms such as RF, support vector machines, and artificial neural networks have demonstrated proficiency in predicting pharmacokinetic parameters like drug absorption and bioavailability.[48,49] Ye et al[50] leveraged transfer learning and multitask learning on a dataset exceeding 30 million bioactivity points to estimate human pharmacokinetic parameters such as oral bioavailability, plasma protein binding, and volume of distribution (Vd). Their deep learning model achieved a mean absolute error (MAE) of 0.31 for oral bioavailability and 0.17 for Vd, outperforming traditional ML techniques.[50] Lou et al[51] extended this capability by focusing on monoclonal antibodies (mAbs). They utilized sequence and structure-based features, such as isoelectric point, solubility

score, and aggregation propensity, to construct a classification model predicting subcutaneous bioavailability. Adaptive boosting and tree-based ML methods achieved an accuracy of 78 %.[51] These advancements underscore the transformative role of AI in accurately modeling absorption parameters.

Similarly, ML and DL techniques have significantly contributed to understanding the distribution of drug candidates within biological systems. For example, Obrezanova et al. employed multitask convolutional neural networks to compute pharmacokinetic profiles and in vivo parameters such as clearance and volume of distribution based on molecular structures and ADME data of over 3000 compounds.[52] These methodologies facilitate the exploration of molecular design space while minimizing resource-intensive experimental assays. Metabolism modeling often requires insights into how chemical structures influence enzymatic reactions. ML techniques, including gradient boosting machines (GBMs) and deep neural networks, have proven effective in predicting metabolic stability and intrinsic clearance. The combination of matched molecular pairs (MMPs) with DNNs has demonstrated superior performance in bioactivity prediction, surpassing traditional ML techniques such as RF and GBMs.[53,54] Furthermore, Koutsoukas et al. highlighted the success of multilayer perceptron models (MLPs) in predicting pKi and pIC_{50} values for multiple targets, showcasing their applicability to metabolic pathway predictions.[55]

AI has also been employed in the prediction of excretion profiles. Quantitative structure–activity relationship (QSAR) models, coupled with advanced DL frameworks like convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, provide robust predictions for drug clearance and half-life.[48,49] These tools enhance lead compound optimization by predicting drug retention times, minimizing risks associated with undesirable pharmacokinetics. Toxicological evaluation remains a pivotal aspect of ADMET prediction. AI-driven models such as DeepTox have demonstrated high accuracy in identifying toxicological risks associated with novel compounds.[56] Open-source platforms like TargeTox and PrOCTOR further empower researchers to address toxicity concerns, enabling efficient identification and mitigation of adverse effects.[57,58] Additionally, interpolation algorithms that leverage structural similarities have been employed to assess environmental and human toxicity risks, thereby ensuring comprehensive safety profiling. In essence, the integration of predictive tools for ADMET attributes significantly enhances lead optimization workflows. For instance, admetSAR 2.0, a computational platform incorporating advanced QSAR modules, facilitates the prediction of multiple pharmacokinetic and pharmacodynamic properties simultaneously. This capability aids researchers in refining lead compounds by balancing efficacy, safety, and pharmacokinetics.[59] The use of AI in modeling drug-receptor interactions further extends to predicting drug efficacy and toxicity, contributing to precision medicine approaches.[60,61]

QSAR modeling has emerged as a pivotal computational technique in the lead optimization phase of drug discovery. It serves as a bridge to predict the biological activity of compounds based on their chemical structure and physicochemical properties. Through the integration of mathematical and statistical models, QSAR facilitates the identification of lead compounds by predicting their activity, toxicity, and other pharmacological properties. This capability is particularly valuable in streamlining the selection process and minimizing experimental costs. The foundation of QSAR was laid in 1964 with the development of the Hansch equation, which correlated physicochemical descriptors such as hydrophobic, electronic, and steric parameters with biological activity through linear regression.[62] Concurrently, the Free-Wilson method was introduced, allowing direct prediction of biological activity from chemical structures without reliance on physicochemical criteria.[63] These methodologies marked the advent of structure-activity relationship analysis in medicinal chemistry.

The evolution of QSAR has been propelled by advancements in machine learning and deep learning, enabling the development of more sophisticated mathematical paradigms. Modern ML algorithms such as random forests, support vector machines, and artificial neural networks have been effectively employed for QSAR modeling.[64,65] Lately, DL has revolutionized QSAR by automating feature extraction and enabling the processing of complex chemical attributes, thus improving prediction accuracy and robustness.[66–68] QSAR modeling has significantly benefited from the integration of ML and DL. For instance, Dahl et al. demonstrated the superiority of multitask deep neural networks over single-task networks in recognizing shared features across related tasks.[66] By leveraging multi-task DNN within platforms like DeepChem, researchers have shown enhanced performance in drug advancement.[68] Moreover, Ma's team, winners of the 2012 Merck Molecular Activity Challenge, highlighted the efficacy of ensemble models combining DNN, Gaussian process regression, and gradient boosting machines for QSAR tasks.[69] Such approaches have proven instrumental in predicting the binding affinities of specific targets, exemplified by the human β-secretase 1 (hBACE-1) blockers, where DNN coupled with 2D molecular descriptors outperformed traditional force-field-dependent techniques.[50]

QSAR has played a critical role in predicting the toxicity and safety profiles of drug candidates, addressing one of the most challenging aspects of drug development. For instance, QSAR models have been employed to predict various toxicity endpoints, including organ toxicity, genotoxicity, and cytotoxicity (Table 1).[70,71] Modern QSAR techniques, incorporating ML and DL, have further improved the accuracy of these predictions. Williams et al. reported the successful application of ML to predict drug-induced liver injury (DILI) with high specificity and sensitivity.[72] Similarly, Verma and Matthews[73] demonstrated the efficacy of combining quantitative structure–toxicity relationship (QSTR) models with ANN for predicting eye and skin irritations, achieving accuracies exceeding 95 %.[73] The continuous refinement of QSAR methodologies has made it an indispensable tool in the hit-to-lead refinement phase. Recent developments in DL-aided QSAR algorithms, such as the use of Canvas descriptors and DNN, have shown remarkable predictive performance in both classification and regression tasks, exemplified by their application to hBACE-1 inhibitors.[67] Additionally, the integration of transcriptional and molecular profiling data with QSAR has enabled the prediction of cardiotoxicity and other adverse effects, advancing the paradigm of precision medicine.[74]

Molecular docking is a cornerstone of structure-based drug design, providing insights into the binding conformation and affinities of small molecules with target proteins. While traditional docking techniques rely on scoring functions to predict ligand-receptor interactions, their limitations in accuracy and computational efficiency have paved the way for integrating machine learning. ML approaches enhance docking by refining scoring functions, identifying binding conformations, and predicting affinities, thereby im-

proving the accuracy and efficiency of virtual screening and lead optimization. ML-driven molecular docking has been instrumental in identifying potent inhibitors of key proteins implicated in neurodegenerative diseases (NDDs). For instance, a study applied virtual screening combined with ML and molecular docking to identify class I and IIb histone deacetylase (HDAC) inhibitors, which are known to promote neurotoxicity in Alzheimer's disease (AD).[75] Here, ML algorithms were employed post-VS to classify potential inhibitors and non-inhibitors, showcasing the utility of ML in narrowing down high-quality hits. Descriptors derived from molecular dynamics (MD) simulations have also been used to train artificial neural networks (ANN) and random forest models for predicting caspase-8 inhibitors. Caspase-8 is a protease implicated in AD pathogenesis, and this integration of ML with docking techniques facilitated the identification of potential leads.[76]

Similarly, ML and docking have been combined to explore traditional Chinese medicine databases for inhibitors targeting glycogen synthase kinase 3 beta (GSK3β), a critical enzyme in AD. Deep learning (DL) was employed for activity prediction, followed by MD simulations to assess the stability of GSK3β-ligand interactions, highlighting the role of ML in predicting bioactivity and structural stability.[77] QSAR modeling has complemented ML in docking-based studies. For example, a QSAR model built with ML algorithms and hybridization techniques identified inhibitors of β-secretase 1 (BACE1), a protein responsible for amyloid-β aggregation in AD. This study utilized a combination of backward elimination and visual analysis to select molecular descriptors, underscoring the precision achievable through ML-driven QSAR modeling.[78]

In Parkinson's disease (PD), cascading Naïve Bayes networks identified potent inhibitors of c-Abl, a kinase promoting neuroprotection. This approach highlights how ML algorithms can refine the docking process to discover not only potent but also safe therapeutic agents.[79] The incorporation of MD simulations further enhances ML-augmented docking by evaluating the dynamic stability of ligand-protein interactions. For instance, Shao et al. combined support vector machine (SVM) algorithms with Tanimoto similarity clustering to identify dual antagonists for the A2A adenosine and dopamine D2 receptors (Table 1). This approach, followed by in vitro experiments, demonstrated neuroprotective potential in PD.[80] Other examples include the use of ML, docking, and MD simulations to identify inhibitors for NLRP3 inflammasome, implicated in PD pathogenesis, using AI-QSAR models and docking for activity prediction and stability assessment.[81]

As well as Galectin-3, a protein associated with neuroinflammation in Huntington's disease (HD), where ML models predicted bioactivity and MD simulations validated binding stability.[82] ML-augmented docking has also facilitated drug repurposing for NDDs. A notable example is the development of deepDR, a deep learning-based drug repurposing tool for AD and PD, which integrates docking and ML for predicting novel therapeutic uses of existing drugs.[83] Another study proposed telmisartan as a potential AD therapy using genetic network-driven ML models, showcasing the versatility of ML in both de novo drug discovery and repurposing efforts.[84]

2.4. Drug repurposing

Drug repurposing involves identifying new therapeutic applications for existing drugs, offering significant advantages in terms of time and cost savings. This strategy bypasses the early stages of drug development, such as preclinical trials, and directly progresses to advanced clinical testing. Notably, the financial burden associated with drug repurposing is substantially lower, with costs estimated at approximately USD 8.4 million compared to USD 41.3 million for developing a new drug entity.[85,86] The integration of machine learning and artificial intelligence has revolutionized this field by enabling the analysis of large, diverse datasets and facilitating the discovery of novel drug-disease associations.

ML techniques, including Support Vector Machines, Neural Networks, and Deep Learning, have played pivotal roles in advancing drug repurposing efforts. SVM, a supervised learning technique, classifies drug-target interactions by constructing hyperplanes that separate data categories with the maximum margin. This approach is particularly effective in network-based drug repurposing, where it analyzes similarities among drugs, diseases, and biological networks.[87] For instance, SVM models have been employed to predict therapeutic targets using multi-dimensional datasets comprising genomic, chemical, and protein interaction information. These models' ability to handle high-dimensional data makes them valuable for extracting insights into drug efficacy and disease mechanisms.

Neural Networks, inspired by the human brain, excel in capturing complex, non-linear relationships within large datasets. Their application in drug repurposing includes predicting multi-target interactions and integrating data sources such as genomics, proteomics, and chemical structures.[88,89] By modeling heterogeneity in drug-disease associations, NNs enable the identification of drugs with potential off-label uses. For instance, they have been instrumental in analyzing patient-specific biomarkers and integrating molecular docking data with biological pathways to uncover novel therapeutic applications. Deep Learning, a subset of NNs, has further transformed drug repurposing by employing hierarchical data modeling to uncover intricate patterns. Tools like DeepDR utilize DL to integrate multiple biological networks, such as drug-disease and drug-target networks, predicting candidates for drug repositioning. Examples include the identification of risperidone and aripiprazole for Alzheimer's disease and methylphenidate and pergolide for Parkinson's disease.[83] DL models such as DTI-RCNN, which combine recurrent neural networks with drug-target interaction data, have achieved high predictive accuracy for therapeutic agents targeting fibrosis and cancer.[90]

Specialized algorithms like PREDICT and SPACE have also made significant contributions to drug repurposing. These logistic regression-based models analyze drug-disease associations by integrating chemical properties, gene expression profiles, and known relationships.[91] PREDICT has demonstrated efficacy in prioritizing drug repurposing opportunities for diseases with limited treatment options. Similarly, SPACE incorporates high-dimensional data such as gene regulatory and protein-protein interaction networks, enabling the identification of high-confidence therapeutic candidates. Network-based methods, such as DTINet, have further revolutionized the field by modeling complex relationships among drugs, targets, and diseases. For example, DTINet integrates multiple

biological networks using dimensionality reduction and network diffusion algorithms, predicting novel cyclooxygenase inhibitors like alendronate and chlorpropamide, which were experimentally validated for their anti-inflammatory effects.[92]

AI-driven drug repurposing has demonstrated remarkable success in addressing critical medical challenges, particularly during the COVID-19 pandemic. Network-based models identified 16 potential anti-HCoV drugs, while multimodal deep learning approaches proposed 12 promising therapeutic targets.[93,94] Similarly, algorithms like RepCOOL identified doxorubicin and paclitaxel as therapeutic agents for stage II breast cancer.[95] Furthermore, the integration of ML with molecular docking has enhanced drug repurposing workflows. Techniques like homology modeling address challenges in molecular docking by predicting interactions for targets lacking crystal structures, thereby expanding the scope of repurposing frameworks.[96]

The incorporation of Real-World Data (RWD) from clinical settings, electronic health records (EHR), patient registries, and wearable devices has emerged as a crucial approach in drug repurposing. RWD offers invaluable insights into the safety, efficacy, and real-world applicability of drugs, complementing traditional clinical trials. Unlike controlled trial environments, RWD captures variability across diverse populations and scenarios, addressing the limitations of homogenous trial cohorts. By leveraging patient-specific data on comorbidities, medication adherence, and genetic predispositions, ML models can identify drug repurposing opportunities for complex diseases like diabetes and neurodegenerative disorders.[97–99]

Network-based approaches integrated with RWD further uncover intricate relationships among drugs, targets, and diseases. For example, drug-disease networks constructed from EHR and genomic data enable the identification of shared pathways and molecular mechanisms. Tools like DTINet utilize RWD to harmonize diverse biological networks, facilitating the prediction of novel drug-target interactions.[92] Predictive models such as PREDICT and SPACE have been enhanced through the integration of patient-centric datasets, including longitudinal health records and pharmacovigilance reports, enabling the identification of high-confidence drug candidates.[91] The application of RWD has led to several notable successes, such as the identification of potential antiviral candidates during the COVID-19 pandemic and repurposing opportunities for chronic conditions like fibrosis and cancer.[90,95]

3. Advances in Ml methodologies for drug discovery

The area of drug discovery has seen substantial progress with the introduction of machine learning (ML) techniques in recent years. Developments in computational science have sped up both the discovery and development of drugs. Traditionally, drug discovery relied on a comprehensive approach emphasizing complete treatment.[102,103] In the last century, healthcare communities worldwide began adopting an allopathic method for treatment and healing.[104] Various computer-based strategies have been employed for the identification and refinement of lead compounds, such as molecular docking, pharmacophore modeling, decision forests, and comparative molecular field analysis.[105] Machine learning and deep learning (DL) have emerged as compelling strategies for drug discovery. The use of ML and DL algorithms spans the entirety of the drug discovery process, rather than being restricted to a singular phase.[106] The capacity for new data analytics to integrate with classical methods and existing hypotheses to generate innovative ones has proven beneficial in areas such as repositioning, target identification, small molecule discovery, and synthesis, among others.[107,108] The data produced within the medical and multi-omic domains is complex and multidimensional. By applying ML techniques, such as generalized linear models via Naive Bayes (NB), the challenges related to analyzing and interpreting multidimensional data may be alleviated.[109] Additional ML methodologies and models frequently utilized in these analyses encompass regression, clustering, regularization, neural networks (NNs), decision trees, dimensionality reduction, ensemble techniques, rule-based systems, and instance-based approaches.[107,110]

3.1. Deep learning (DL) architectures: CNNs, RNNs, and attention-based models tailored for molecular data

Deep learning (DL) algorithms are regarded as one of the leading areas of advancement and research across nearly all scientific and technological domains.[111] The resurgence of artificial neural networks (NNs) into practical algorithms from their original theoretical and anticipated uses, first initiated in the 1950s, serves as a crucial foundation for DL and the ongoing success attributed to AI-based integration of standard methods.[112] DL algorithms empower computational models to learn representations of multidimensional data through abstraction.[113] DL has made significant strides in addressing numerous challenges encountered by traditional machine learning algorithms, such as image and speech recognition.[114] Within the drug discovery landscape, DL methods have emerged as exemplary techniques for predicting drug activity, discovering targets, and identifying lead molecules.[115] The foundation of DL is frequently linked to NN systems, which are employed to develop systems capable of executing complex data recognition, interpretation, and creation. The principal subsets of artificial NNs currently utilized in drug discovery include deep neural networks, recurrent neural networks, and convolutional neural networks.[116]

DL algorithms function through neural networks that can integrate other ML techniques for training. By employing supervised and reinforcement learning-guided strategies, a DNN can be trained to perform intricate tasks.[107,117] A generative DNN can synthesize novel chemical compounds based on existing libraries and training datasets**.** Meanwhile, a predictive DNN can forecast the chemical properties of these novel compounds.[118,119]

Quantitative Structure-Activity Relationship models are currently utilized to discover the correlation between the chemical structure of these compounds and their activity. QSAR analysis represents one of the most sophisticated forms of DL-based AI in contemporary drug discovery and development.[120] It has enabled researchers to transform 2D chemical structures into physicochemical descriptors associated with the activity of the molecule. The advancement of 3D-QSAR has facilitated deeper investigations into how geometric structure influences ligand-target interactions.[121] QSAR has been actively employed in the pharmaceutical sector to predict both on-target and off-target activities of developed lead compounds on specific targets.

3.2. Natural language processing (NLP) for text mining

Natural Language Processing (NLP) techniques have greatly progressed the area of biomedical literature mining, offering powerful methods for extracting and interpreting intricate textual information.[122] Essential NLP methods encompass Named Entity Recognition (NER), Relation Extraction, and Text Classification, all of which are vital for processing biomedical texts.[123,124] Named Entity Recognition (NER) entails the detection and categorization of specific entities within a text, including genes, proteins, diseases, and chemicals. This function is crucial for comprehending biomedical literature, as it facilitates the extraction of pertinent entities from unstructured text.[125] Sophisticated NER models utilize pre-trained embeddings and contextual clues to precisely identify entities, even amid ambiguous or novel terminology. For example, specialized NER systems like Biocreative and those based on the NCBI disease corpus are specifically designed to identify entities within the biomedical field, thereby enhancing the accuracy of entity extraction.[126]

A primary function of NLP in the realm of drug discovery is the automated extraction and integration of significant data from scientific publications. Conventional data extraction techniques are frequently labor-intensive and slow, while NLP models simplify this task by effectively recognizing and indexing crucial information such as drug interactions, side effects, and biomarker correlations.[127] For example, NLP algorithms can quickly analyze and categorize large sets of literature to identify emerging patterns and insights that guide drug development efforts.[128]

Additionally, NLP aids in the discovery of novel drug targets by examining patterns and connections within biomedical texts. By extracting data on protein-protein interactions, gene expression patterns, and disease mechanisms, NLP tools can propose new targets for drug development that might not be obvious through traditional approaches.[129] This functionality accelerates the early phases of drug discovery, where pinpointing promising targets is vital for further research. NLP also plays a role in enhancing drug repurposing initiatives by analyzing existing literature and clinical trial information, NLP models can reveal potential new applications for already established drugs. This strategy utilizes existing safety and efficacy data to suggest alternative therapeutic uses, thereby decreasing the time and expenses related to developing entirely new drugs.[130,131]

Identifying therapeutic targets and understanding mechanisms of action are essential aspects of drug discovery, and sophisticated NLP models are integral to these processes.[132] These models utilize deep learning and contextual embeddings to evaluate intricate biomedical data and provide actionable insights. Therapeutic target identification entails finding molecules or pathways that significantly influence disease progression and can be influenced by therapeutic agents.[133] NLP models, including those utilizing transformer architectures like BERT or GPT, are used to analyze extensive omics data and scientific literature to identify prospective targets.[134,135] These models examine relationships among genes, proteins, and diseases, identifying targets most likely to offer therapeutic advantages. For instance, by combining data from genomic research and literature, NLP models can spotlight genes with elevated expression levels in certain cancers, indicating them as potential drug targets.[136,137] Comprehending these mechanisms is crucial for developing effective drugs and predicting their side effects.

NLP models assist in clarifying mechanisms of action by extracting and synthesizing data from experimental findings, clinical reports, and drug databases.[138] By investigating how drugs interact with biological targets and their subsequent effects, NLP tools offer insights into the pathways and processes influenced by the drug, which helps in designing more precise and effective therapies.[139]

Integrating biomedical literature with drug discovery databases is essential for utilizing NLP tools to improve drug development. This combination allows for the synthesis of information from various sources, offering a complete perspective on the connections between drugs, targets, and diseases.[140] Biomedical literature is an abundant source of knowledge regarding the effects, side effects, and interactions of drugs.[141] Merging this literature with drug discovery databases enables researchers to better comprehend existing medications and their possible uses. For example, combining literature data with databases like DrugBank or PubChem helps identify new drug candidates and investigate their chemical characteristics and biological functions.[142,143]

3.3. Transfer learning and few-shot learning

Transfer learning is a machine learning strategy where a model developed for one task is modified or utilized for a related task. This approach entails pretraining a model on a broad, general dataset, followed by fine-tuning it on a smaller, domain-specific dataset. This technique proves especially beneficial in drug discovery, particularly when labeled data is scarce.[144–146] The majority of deep transfer learning research has focused on predicting compound properties, generating molecules, and performing structure-based virtual screening.

The multi-channel PINN a neural network model designed for assessing compound–target protein interactions, was trained to classify toxicity (activity) as the target task for transfer learning. When evaluated, transfer learning demonstrated greater success compared to training a model from the ground up.[147,148] Another detailed investigation into transfer learning in DTI prediction (chemogenomic DTI binary classifier) was conducted by Playe and Stoven.[149] They indicated that utilizing the full fine-tuning approach for transfer learning could enhance prediction performance when the source task closely resembles the target task. Dey and colleagues explored instance-based and feature-based transfer learning, contrasting with the more commonly used parameter-based transfer learning, like pre-training.[150] Yang et al. applied transfer learning to forecast protein–protein interactions.[151]

Generative models that have been pretrained, such as those based on Transformer or Recurrent Neural Network (RNN) architectures, are modified to design molecules with defined properties. These models are educated using SMILES representations from expansive chemical libraries, generating molecules aimed at specific receptors like GPCRs.[152] Models that have been pretrained on datasets concerning gene expression or proteomics can predict interactions between drugs and genes for emerging diseases. Training occurs on datasets that reflect cancer gene expression, predicting drug effectiveness for infrequent diseases.[153] Few-shot learning

(FSL) empowers models to perform effectively on tasks with a minimal number of labeled examples. This is particularly significant in drug discovery for rare diseases or new targets, where data availability is limited.[154] Few-shot models derive insights from a small number of known drug-target associations to forecast interactions for unfamiliar drugs or targets. They can predict new uses for drugs based on 10–20 examples of recognized drug-disease links.[142]

FSL can also be employed to model the efficacy of drugs targeting rare diseases using limited patient data. By utilizing patient omics profiles, these models can be trained on a small group to anticipate drug responses.[155] Few-shot models are capable of predicting potential protein targets for new small molecules with scant data, learning from a few known ligands for an unseen protein and generalizing to new ligands.[156] Few-shot learning is also applied to estimate bioactivity (e.g., IC_{50}, EC_{50}) from a small number of experimental outcomes, where the model learns from a limited selection of labeled assays for a new compound.[157]

3.4. Federated learning

Federated learning (FL) represents a machine learning approach that permits various organizations to work together on model training without disclosing sensitive information.[158] Rather than consolidating raw data on a centralized server, FL conducts model training locally using each organization's data and only shares model updates (like gradients or weights), thereby safeguarding privacy.[159] In the field of drug discovery, FL promotes cooperation among pharmaceutical firms, medical facilities, and research organizations while adhering to privacy laws and regulations such as GDPR or HIPAA.[160]

Drug discovery frequently necessitates the examination of patient omics data (including genomics and proteomics) that is dispersed across multiple organizations. The sharing of such data is often limited by privacy restrictions.[161] FL enables organizations to collaboratively develop models to pinpoint disease-specific drug targets without transmitting sensitive patient information.[162] For instance, hospitals can train a unified model to discover biomarkers for Alzheimer's disease utilizing their local genomic datasets. Model updates are combined without revealing individual patient details.[163]

Knowledge regarding drug responses is spread across various pharmaceutical firms and healthcare providers. FL permits organizations to jointly train models on patient response data (such as electronic health records and clinical trials) to uncover new uses for existing medications.[164] FL assists in evaluating antiviral drug responses among diverse populations against diseases like COVID-19.[165]

Gaining insight into drug-drug interactions (DDIs) requires diverse datasets sourced from multiple repositories, including electronic health records, clinical trials, and pharmacological databases.[166] FL can train a global model for predicting DDIs based on these distributed datasets while maintaining data confidentiality. FL enables these companies to work together to train models for predicting toxicity without revealing proprietary compound information.[167,168] Acquiring real-world data (such as patient outcomes) across organizations is often limited due to privacy issues. FL allows institutions to derive insights from real-world evidence while keeping patient information private.[169] Numerous pharmaceutical firms aim to estimate compound toxicity while safeguarding their proprietary chemical libraries. Each company independently trains the model using its own chemical toxicity data. The model updates are combined through FedAvg as a result, and a shared model emerges that accurately predicts compound toxicity by leveraging a variety of datasets without risking proprietary information.[170]

4. Challenges and limitation

The use of machine learning (ML) in the discovery of pharmaceutical drugs presents significant potential for transformation, but it also encounters numerous challenges and limitations, including those related to data, technical aspects, and regulatory hurdles.

Pharmaceutical datasets often exhibit inconsistencies, sparsity, or bias, creating notable barriers to the training and generalization of ML models.[171] Information sourced from various areas such as clinical trials, genomics, proteomics, and electronic health records is frequently stored in incompatible formats and measured against different standards. This lack of standardization complicates the integration of datasets, which is essential for developing robust ML models.[172]

Datasets often have missing values, especially during the initial phases of drug discovery, where experimental evidence might be incomplete or unreliable. For rare diseases or innovative drug targets, the limited size of datasets can hinder the training of complex ML models, diminishing their predictive power.[173] Access to high-quality datasets is restricted by stringent regulations and intellectual property laws, further limiting the data availability needed for model training. Measurement errors and variability in experiments can introduce noise into the datasets, compromising the reliability of ML predictions.[174]

Advanced ML models, particularly deep learning techniques like graph neural networks (GNNs) and transformers, excel in feature extraction but lack transparency. Their predictions related to drug efficacy, toxicity, or safety are challenging to interpret, which can undermine the confidence of regulatory agencies and researchers.[175] Also, Regulatory bodies require clear explanations for ML-driven predictions, especially for critical life-dependent decisions. The "black-box" characteristic of many models complicates the path to regulatory approval.[176] Legislation such as GDPR and HIPAA restricts the sharing and use of sensitive patient information, thereby limiting access to critical training datasets.[177] There is a lack of standardized procedures for validating ML models in drug discovery, leading to delays in regulatory approval. Establishing accountability when ML-generated predictions result in negative outcomes during clinical trials adds another layer of complexity.

5. Future directions

Overcoming these challenges requires a cohesive and forward-thinking approach that combines innovation in AI, collaboration across disciplines, and the integration of diverse data sources. A crucial step involves developing interpretable AI methodologies capable of providing clear reasoning for predictions, such as why a specific drug molecule is likely to be effective or toxic. This is especially critical in clinical trials, where regulatory authorities and clinicians demand transparency and a deep understanding of model predictions. Techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are instrumental in identifying molecular or biological features driving predictions. Complementing these approaches, tools designed to visualize molecular interactions, decision boundaries, or model outputs in a user-friendly manner can empower chemists and biologists, bridging the gap between advanced AI and practical scientific application.

Integrating mechanistic models, such as pharmacokinetic simulations, with machine learning can further provide scientifically grounded and interpretable insights. Leveraging diverse data modalities, including genomics, proteomics, transcriptomics, and clinical records, enhances the comprehensiveness of drug discovery pipelines. Graph neural networks (GNNs) emerge as a powerful tool for representing and integrating complex relationships across data types, such as drug-target, protein–protein, and gene-disease interactions. Similarly, methods like variational autoencoders (VAEs) can harmonize disparate data types into a unified latent space, improving both integration and interpretability. Advocating for the adoption of data standards, like the FAIR principles (Findable, Accessible, Interoperable, and Reusable), ensures that datasets across the pharmaceutical industry are compatible and usable.

The integration of machine learning with traditional methods, such as QSAR (Quantitative Structure-Activity Relationship) modeling, enables the combination of computational power and expert-driven insights. Building ML models that incorporate biological mechanisms and pathway knowledge provides more accurate predictions of drug efficacy and safety. Active learning techniques, which refine models by prioritizing experiments based on predicted uncertainties, enhance iterative improvement. Similarly, combining atomistic simulations like molecular dynamics with ML models provides detailed insights into drug interactions. Establishing feedback loops where ML predictions are experimentally validated and used to refine models ensures a dynamic and improving system.

Collaboration is essential to overcoming the data silos and resource constraints inherent in the pharmaceutical industry. Promoting partnerships between pharmaceutical companies, academic institutions, and technology providers fosters the sharing of expertise and resources. Pre-competitive consortia focusing on foundational challenges, such as data standardization and validation, can drive progress. Federated learning (FL) enables collaborative model-building across institutions while preserving data privacy. Shared repositories of preclinical and clinical datasets, such as ChEMBL and PDB, along with pre-trained ML models, democratize access to crucial resources. Unified protocols for data collection, annotation, and reporting enhance compatibility and reproducibility across organizations.

Optimizing drug candidates involves iterative improvements to molecular properties or synthetic routes. The integration of AI with robotic lab systems accelerates experimental validation, while quantum computing holds potential for simulating molecular interactions and solving optimization problems with unprecedented speed. These advancements promise to not only accelerate the drug discovery process but also make it more precise, interpretable, and collaborative, ultimately transforming the way therapeutics are developed and delivered.

6. Conclusion

In recent years, machine learning has emerged as a transformative force in drug discovery, addressing longstanding challenges in drug development through innovative methodologies. Key advances include the application of deep learning models, such as convolutional neural networks, recurrent neural networks, and transformer-based architectures, which have significantly improved molecular property prediction, ligand-target interaction analysis, and protein structure modeling. Natural language processing tools like SciBERT and BioBERT have enabled the extraction of actionable insights from vast biomedical literature, accelerating hypothesis generation and drug repurposing opportunities. Moreover, the integration of transfer learning and few-shot learning approaches has proven invaluable for tasks with limited data, such as predicting toxicity profiles and optimizing lead compounds, while federated learning has facilitated secure multi-institutional collaborations without compromising data privacy. The potential of ML to transform drug discovery is vast. By improving the speed and accuracy of lead compound identification, optimizing clinical trial design, and integrating multi-omics datasets, ML methodologies promise to streamline the entire drug discovery pipeline. These innovations hold the key to reducing development costs, increasing the probability of clinical success, and ultimately delivering new therapies to patients faster than ever before. Particularly in the era of precision medicine, ML-driven approaches can help tailor treatments to individual patients based on molecular profiles and real-world data, ushering in a new generation of targeted therapeutics. Despite these advances, the path forward requires sustained efforts in several critical areas. Continued development of robust, interpretable ML models is essential for ensuring reliable predictions and minimizing biases. Ethical considerations, such as ensuring patient data privacy and addressing algorithmic fairness, must remain central to these developments. Furthermore, successful implementation will depend on interdisciplinary collaboration among computational scientists, biologists, clinicians, and regulatory bodies. Only through such concerted efforts can the full potential of ML be realized, paving the way for a new paradigm in drug discovery that is faster, more cost-effective, and ultimately life-changing for patients worldwide.

Funding

There is no funding to report.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Authors Contribution

UEO: Conceptualization, Writing- Editing & Review. **EFO, OOB, OOP, SO, TEG, SB, CAO and JOO**: Writing- Original Draft, Writing- Editing & Review. **CGC**: Writing- Editing & Review, Software, Resources. **CCA:** Writing- Editing & Review, Visualisation, Resources, Writing- Editing & Review. **AF:** Writing- Editing & Review Project Administration. **HO:** Conceptualization, Writing- Editing & Review, Software, Supervision.

Ethics approval and consent to participate

Not applicable.

Conflict of Interest

The authors declare that there is no known conflict of interest regarding this paper publication.

Acknowledgments

None.

References

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ*. 2016;47:20–33.
2. Waring MJ, Arrowsmith J, Leach AR, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discov*. 2015;14(7):475–486.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
4. Schneider P, Walters WP, Plowright AT, et al. Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov*. 2020;19(5):353–364.
5. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23(6):1241–1250.
6. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater*. 2019;18(5):435–441.
7. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37(9):1038–1040.
8. Ye L, Ngan DK, Xu T, et al. Prediction of drug-induced liver injury and cardiotoxicity using chemical structure and in vitro assay data. *Toxicol Appl Pharmacol*. 2022;454:116250.
9. Cavasotto CN, Scardino V. Machine learning toxicity prediction: latest advances by toxicity end point. *ACS Omega*. 2022;7(51):47536–47546.
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
11. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open targets platform: new developments and updates two years on. *Nucleic Acids Res*. 2019;47(D1):D1056–D1065.
12. Costa PR, Acencio ML, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genom*. 2010;11:1–15 BioMed Central.
13. Jeon J, Nim S, Teyra J, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Med*. 2014;6:1–18.
14. Mamoshina P, Volosnikova M, Ozerov IV, et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front Genet*. 2018;9:242.
15. Li B, Shin H, Gulbekyan G, et al. Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. *PLoS One*. 2015;10(6):e0130700.
16. van Gool AJ, Bietrix F, Caldenhoven E, et al. Bridging the translational innovation gap through good biomarker practice. *Nat Rev Drug Discov*. 2017;16(9):587–588.
17. Kraus VB. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat Rev Rheumatol*. 2018;14(6):354–362.
18. Shi L, Campbell G, Jones WD. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010;28(8):827–838.
19. Zhan F, Huang Y, Colla S, et al. The molecular classification of multiple myeloma. *Blood*. 2006;108(6):2020–2028.
20. Zhan F, Barlogie B, Mulligan G, Shaughnessy Jr JD, Bryant B. High-risk myeloma: a gene expression–based risk-stratification model for newly diagnosed multiple myeloma treated with high-dose therapy is predictive of outcome in relapsed disease treated with single-agent bortezomib or high-dose dexamethasone. *Blood J Am Soc Hematol*. 2008;111(2):968–969.
21. Shaughnessy Jr JD, Zhan F, Burington BE, et al. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood*. 2007;109(6):2276–2284.
22. Mulligan G, Mitsiades C, Bryant B, et al. Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood*. 2007;109(8):3177–3188.
23. Decaux O, Lodé L, Magrangeas F, et al. Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myelome. *J Clin Oncol*. 2008;26(29):4798–4805.
24. Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*. 2016;166(3):740–754.
25. Ament SA, Pearl JR, Cantle JP, et al. Transcriptional regulatory networks underlying gene expression changes in Huntington's disease. *Mol Syst Biol*. 2018;14(3):e7435.

26. Ferrero E, Dunham I, Sanseau P. In silico prediction of novel therapeutic targets using gene–disease association data. *J Transl Med*. 2017;15:1–16.
27. Rouillard AD, Hurle MR, Agarwal P. Systematic interrogation of diverse Omic data reveals interpretable, robust, and generalizable transcriptomic features of clinically successful therapeutic targets. *PLoS Comput Biol*. 2018;14(5):e1006142.
28. Jha A, Gazzara MR, Barash Y. Integrative deep models for alternative splicing. *Bioinformatics*. 2017;33(14):i274–i282.
29. Vaquero-Garcia J, Barrera A, Gazzara MR, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*. 2016;5:e11752.
30. Rohacek AM, Bebee TW, Tilton RK, et al. ESRP1 mutations cause hearing loss due to defects in alternative splicing that disrupt cochlear development. *Dev Cell*. 2017;43(3):318–331.
31. Valentini G, Paccanaro A, Caniza H, Romero AE, Re M. An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artif Intell Med*. 2014;61(2):63–78.
32. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput Chem*. 2001;26(1):5–14.
33. Vapnik VN. An overview of statistical learning theory. *IEEE Trans Neural Netw*. 1999;10(5):988–999.
34. Saeh JC, Lyne PD, Takasaki BK, Cosgrove DA. Lead hopping using SVM and 3D pharmacophore fingerprints. *J Chem Inf Model*. 2005;45(4):1122–1133.
35. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, Schneider G. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J Med Chem*. 2005;48(22):6997–7004.
36. Wang C, Zhang Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J Comput Chem*. 2017;38(3):169–177.
37. Eitrich T, Kless A, Druska C, Meyer W, Grotendorst J. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J Chem Inf Model*. 2007;47(1):92–103.
38. Lepp Z, Kinoshita T, Chuman H. Screening for new antidepressant leads of multiple activities by support vector machines. *J Chem Inf Model*. 2006;46(1):158–167.
39. Jorissen RN, Gilson MK. Virtual screening of molecular databases using a support vector machine. *J Chem Inf Model*. 2005;45(3):549–561.
40. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol Pharm*. 2017;14(9):3098–3104.
41. Ahmad S, Bano N, Sharma S, Sakina S, Ahmad N, Raza K. Generative AI in drug designing: current state-of-the-art and perspectives. *Generative AI: Current Trends and Applications*. Singapore: Springer Nature Singapore; 2024:427–463.
42. Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. *Int J Mol Sci*. 2019;20(11):2783.
43. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci*. 2018;4(2):268–276.
44. Segler MH, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. 2018;555(7698):604–610.
45. Li X, Xu Y, Yao H, Lin K. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *J Cheminform*. 2020;12:1–13.
46. Perron Q, Mirguet O, Tajmouati H, et al. Deep generative models for ligand-based de novo design applied to multi-parametric optimization. *J Comput Chem*. 2022;43(10):692–703.
47. Assmann M, Bal M, Craig M, et al. A novel machine learning approach uncovers new and distinctive inhibitors for cyclin-dependent kinase 9. *BioRxiv*. 2020 2020-03.
48. Daoui O, Elkhattabi S, Chtita S, Elkhalabi R, Zgou H, Benjelloun AT. QSAR, molecular docking and ADMET properties in silico studies of novel 4, 5, 6, 7-tetrahydrobenzo [D]-thiazol-2-Yl derivatives derived from dimedone as potent anti-tumor agents through inhibition of C-Met receptor tyrosine kinase. *Heliyon*. 2021;7(7).
49. Bhattamisra SK, Banerjee P, Gupta P, Mayuren J, Patra S, Candasamy M. Artificial intelligence in pharmaceutical and healthcare research. *Big Data Cogn Comput*. 2023;7(1):10.
50. Ye Z, Yang Y, Li X, Cao D, Ouyang D. An integrated transfer learning and multitask learning approach for pharmacokinetic parameter prediction. *Mol Pharm*. 2018;16(2):533–541.
51. Lou H, Hageman MJ. Machine learning attempts for predicting human subcutaneous bioavailability of monoclonal antibodies. *Pharm Res*. 2021;38:451–460.
52. Obrezanova O. Artificial intelligence for compound pharmacokinetics prediction. *Curr Opin Struct Biol*. 2023;79:102546.
53. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme gradient boosting as a method for quantitative structure–activity relationships. *J Chem Inf Model*. 2016;56(12):2353–2360.
54. Pereira JC, Caffarena ER, Dos Santos CN. Boosting docking-based virtual screening with deep learning. *J Chem Inf Model*. 2016;56(12):2495–2506.
55. Koutsoukas A, Monaghan KJ, Li X, Huan J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J Cheminform*. 2017;9:1–13.
56. Jain S, Siramshetty VB, Alves VM, et al. Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. *J Chem Inf Model*. 2021;61(2):653–663.
57. Lysenko A, Sharma A, Boroevich KA, Tsunoda T. An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci Alliance*. 2018(6):1.
58. Basile AO, Yahi A, Tatonetti NP. Artificial intelligence for drug toxicity and safety. *Trends Pharmacol Sci*. 2019;40(9):624–635.
59. Yang H, Lou C, Sun L, et al. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics*. 2019;35(6):1067–1069.
60. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. *Am J Med*. 2019;132(7):795–801.
61. Kim H, Kim E, Lee I, Bae B, Park M, Nam H. Artificial intelligence in drug discovery: a comprehensive review of data-driven and machine learning approaches. *Biotechnol Bioprocess Eng*. 2020;25:895–930.
62. Hansch C, Fujita T. Additions and corrections-ρ-σ-π analysis. a method for the correlation of biological activity and chemical structure. *J Am Chem Soc*. 1964;86(24):5710.
63. Free SM, Wilson JW. A mathematical contribution to structure-activity studies. *J Med Chem*. 1964;7(4):395–399.
64. Ning X, Karypis G. In silico structure-activity-relationship (SAR) models from machine learning: a review. *Drug Dev Res*. 2011;72(2):138–146.
65. Dobchev D, Pillai G, Karelson M. In silico machine learning methods in drug development. *Curr Top Med Chem*. 2014;14(16):1913–1922.
66. Dahl, G.E., Jaitly, N., & Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231*.
67. Subramanian G, Ramsundar B, Pande V, Denny RA. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J Chem Inf Model*. 2016;56(10):1936–1949.
68. Ramsundar B, Liu B, Wu Z, et al. Is multitask deep learning practical for pharma? *J Chem Inf Model*. 2017;57(8):2068–2076.
69. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. *J Chem Inf Model*. 2015;55(2):263–274.
70. Algamal ZY, Lee MH, Al-Fakih AM, Aziz M. High-dimensional QSAR prediction of anticancer potency of imidazo [4, 5-b] pyridine derivatives using adjusted adaptive LASSO. *J Chemom*. 2015;29(10):547–556.
71. Patlewicz G, Fitzpatrick JM. Current and future perspectives on the development, evaluation, and application of in silico approaches for predicting toxicity. *Chem Res Toxicol*. 2016;29(4):438–451.
72. Williams DP, Lazic SE, Foster AJ, Semenova E, Morgan P. Predicting drug-induced liver injury with Bayesian machine learning. *Chem Res Toxicol*. 2019;33(1):239–248.
73. Verma RP, Matthews EJ. Estimation of the chemical-induced eye injury using a weight-of-evidence (WoE) battery of 21 artificial neural network (ANN) c-QSAR models (QSAR-21): part I: irritation potential. *Regul Toxicol Pharmacol*. 2015;71(2):318–330.
74. Mamoshina P, Bueno-Orovio A, Rodriguez B. Dual transcriptomic and molecular machine learning predicts all major clinical forms of drug cardiotoxicity. *Front Pharmacol*. 2020;11:639.

75. Gupta R, Ambasta RK, Kumar P. Identification of novel class I and class IIb histone deacetylase inhibitor for Alzheimer's disease therapeutics. *Life Sci*. 2020;256:117912.
76. Jamal S, Grover A, Grover S. Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against Alzheimer's disease. *Front Pharmacol*. 2019;10:780.
77. Chen HY, Chen JQ, Li JY, et al. Deep learning and random forest approach for finding the optimal traditional chinese medicine formula for treatment of alzheimer's disease. *J Chem Inf Model*. 2019;59(4):1605–1623.
78. Ponzoni I, Sebastián-Pérez V, Martínez MJ, et al. QSAR classification models for predicting the activity of inhibitors of beta-secretase (BACE1) associated with Alzheimer's disease. *Sci Rep*. 2019;9(1):9102.
79. Kaiser TM, Dentmon ZW, Dalloul CE, Sharma SK, Liotta DC. Accelerated discovery of novel ponatinib analogs with improved properties for the treatment of parkinson's disease. *ACS Med Chem Lett*. 2020;11(4):491–496.
80. Shao YM, Ma X, Paira P, et al. Discovery of indolylpiperazinylpyrimidines with dual-target profiles at adenosine A2A and dopamine D2 receptors for Parkinson's disease treatment. *PloS One*. 2018;13(1):e0188212.
81. Chen ZD, Zhao L, Chen HY, Gong JN, Chen X, Chen CYC. A novel artificial intelligence protocol to investigate potential leads for Parkinson's disease. *RSC Adv.*. 2020;10(39):22939–22958.
82. Deng L, Zhong W, Zhao L, et al. Artificial intelligence-based application to explore inhibitors of neurodegenerative diseases. *Front Neurorobot*. 2020;14:617327.
83. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*. 2019;35(24):5191–5198.
84. Oh M, Ahn J, Yoon Y. A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. *PLoS One*. 2014;9(10):e111668.
85. Persidis A. The benefits of drug repositioning. *Drug Discov World*. 2011;12:9–12.
86. Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects. *Drug Discov Today*. 2019;24(3):773–780.
87. Koromina M, Pandi MT, Patrinos GP. Rethinking drug repositioning and development with artificial intelligence, machine learning, and omics. *Omics: J Integr Biol*. 2019;23(11):539–548.
88. Iyengar R, Zhao S, Chung SW, Mager DE, Gallo JM. Merging systems biology with pharmacodynamics. *Sci Transl Med*. 2012;4(126) 126ps7.
89. Wuerth R, Thellung S, Bajetto A, Mazzanti M, Florio T, Barbieri F. Drug-repositioning opportunities for cancer therapy: novel molecular targets for known compounds. *Drug Discov Today*. 2016;21(1):190–199.
90. Wu D, Gao W, Li X, et al. Dr AFC: drug repositioning through anti-fibrosis characteristic. *Br Bioinform*. 2021;22(3):bbaa115.
91. Park K. A review of computational drug repurposing. *Transl Clin Pharmacol*. 2019;27(2):59–63.
92. Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun*. 2017;8(1):573.
93. Hooshmand SA, Zarei Ghobadi M, Hooshmand SE, Azimzadeh Jamalkandi S, Alavi SM, Masoudi-Nejad A. A multimodal deep learning-based drug repurposing approach for treatment of COVID-19. *Mol Divers*. 2021;25:1717–1730.
94. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov*. 2020;6(1):14.
95. Fahimian G, Zahiri J, Arab SS, Sajedi RH. RepCOOL: computational drug repositioning via integrating heterogeneous biological networks. *J Transl Med*. 2020;18:1–10.
96. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem Biol Drug Des*. 2019;93(1):12–20.
97. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. *Br. Bioinform*. 2018;19(5):878–892.
98. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*. 2004;3(8):673–683.
99. Gakire MH. The use of machine learning in predicting disease progression. *Newport IntJ Res Med Sci*. 2024;5(3):1–4. doi:10.59298/nijrms/2024/5.3.14883.
100. Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J Chem Inf Model*. 2018;58(2):287–296.
101. Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. *Bioinformatics*. 2020;36(Supplement_1):i525–i533.
102. Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform*. 2020;12:1–23.
103. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature*. 2023;616(7958):673–685.
104. Salmon JW, ed. *Alternative Medicines: popular and Policy Perspectives*. Taylor & Francis; 2022.
105. Naithani U, Guleria V. Integrative computational approaches for discovery and evaluation of lead compound for drug design. *Front Drug Discov*. 2024;4:1362456.
106. Vijayan RSK, Kihlberg J, Cross JB, Poongavanam V. Enhancing preclinical drug discovery with artificial intelligence. *Drug Discov Today*. 2022;27(4):967–984.
107. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine learning methods in drug discovery. *Molecules*. 2020;25(22):5277.
108. Kiriiri GK, Njogu PM, Mwangi AN. Exploring different approaches to improve the success of drug discovery and development projects: a review. *Fut J Pharm Sci*. 2020;6:1–12.
109. Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and challenges in machine learning: a comprehensive review of models, libraries, applications, and algorithms. *Electronics*. 2023;12(8):1789.
110. Pugliese Viloria ADJ, Folini A, Carrion D, Brovelli MA. Hazard susceptibility mapping with machine and deep learning: a literature review. *Remote Sens*. 2024;16(18):3374.
111. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers*. 2021;25:1315–1360.
112. Worden K, Tsialiamanis G, Cross EJ, Rogers TJ. Artificial neural networks. *Machine Learning in Modeling and Simulation: Methods and Applications*. Cham: Springer International Publishing; 2023:85–119.
113. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci*. 2021;2(6):420.
114. Patil d, Rane NL, Desai P, Rane J. Machine learning and deep learning: Methods, techniques, applications, challenges, and future research opportunities. *Trust Artif Intell Ind Soc*. 2024:28–81.
115. Mak KK, Wong YH, Pichika MR. Artificial intelligence in drug discovery and development. *Drug Discov Eval: Saf Pharmacokinet Assays*. 2024:1461–1498.
116. Vatansever S, Schlessinger A, Wacker D, et al. Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: state-of-the-arts and future directions. *Med Res Rev*. 2021;41(3):1427–1473.
117. Goshisht MK. Machine learning and deep learning in synthetic biology: key architectures, applications, and challenges. *ACS Omega*. 2024;9(9):9921–9945.
118. Sousa T, Correia J, Pereira V, Rocha M. Generative deep learning for targeted compound design. *J Chem Inf Model*. 2021;61(11):5343–5361.
119. Galushka M, Swain C, Browne F, Mulvenna MD, Bond R, Gray D. Prediction of chemical compounds properties using a deep learning model. *Neural Comput Appl*. 2021;33(20):13345–13366.
120. Chakraborty C, Bhattacharya M, Lee SS, Wen ZH, Lo YH. The changing scenario of drug discovery using artificial intelligence (AI) to deep learning (DL): Recent advancement, success stories, collaborations, and challenges. *Mol Ther-Nucleic Acids*. 2024;35(3):102295.
121. Iwaloye O, Ottu PO, Olawale F, et al. Computer-aided drug design in anti-cancer drug discovery: what have we learnt and what is the way forward? *Inform Med Unlocked*. 2023;41:101332.
122. Zhao S, Su C, Lu Z, Wang F. Recent advances in biomedical literature mining. *Brief Bioinform*. 2021;22(3):bbaa057.
123. Durango MC, Torres-Silva EA, Orozco-Duque A. Named entity recognition in electronic health records: a methodological review. *Healthc Inform Res*. 2023;29(4):286.
124. Pakhale, K. (2023). Comprehensive overview of named entity recognition: models, domain-specific applications and challenges. *arXiv preprint arXiv:2309.14084*.

125. Janowski A. Natural language processing techniques for clinical text analysis in healthcare. *J Adv Anal Healthc Manag*. 2023;7(1):51–76.
126. Perera N, Dehmer M, Emmert-Streib F. Named entity recognition and relation detection for biomedical information extraction. *Front Cell Dev Biol*. 2020;8:673.
127. Xu X, Li J, Zhu Z, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. *Bioengineering*. 2024;11(3):219.
128. Bhatnagar R, Sardar S, Beheshti M, Podichetty JT. How can natural language processing help model informed drug development?: a review. *JAMIA Open*. 2022;5(2):ooac043.
129. Mitta NR. Application of AI-driven natural language processing in biomedical literature mining: developing deep learning models for automated knowledge extraction, hypothesis generation, and drug discovery insights. *Afr J Artif Intell Sustain Dev*. 2023;3(2):361–397.
130. Islam MT, Newaz AAH, Paul R, Melon MMH, Hussen M. Ai-driven drug repurposing: uncovering hidden potentials of established medications for rare disease treatment. *Libr Prog Int*. 2024;44(3):21949–21965.
131. Gangwal A, Ansari A, Ahmad I, Azad AK, Sulaiman WMAW. Current strategies to address data scarcity in artificial intelligence-based drug discovery: a comprehensive review. *Comput Biol Med*. 2024;179:108734.
132. Pun FW, Ozerov IV, Zhavoronkov A. AI-powered therapeutic target discovery. *Trends Pharmacol Sci*. 2023;44(9):561–572.
133. Yip HYK, Papa A. Signaling pathways in cancer: therapeutic targets, combinatorial treatments, and new developments. *Cells*. 2021;10(3):659.
134. Liu, J., Yang, M., Yu, Y., Xu, H., Li, K., & Zhou, X. (2024). Large language models in bioinformatics: applications and perspectives. ArXiv.
135. Sarumi OA, Heider D. Large language models and their applications in bioinformatics. *Comput Struct Biotechnol J*. 2024;23:3498–3505.
136. Iyer K, Ivanov J, Tenchov R, et al. Emerging targets and therapeutics in immuno-oncology landscape: insights from natural language processing analysis. *ChemRxiv*. 2023;2023. doi:10.26434/chemrxiv-2023-bpvfq.
137. Greeshma[1] V, Rao CB, Gopal V. The impact of artificial intelligence in gene therapy. *Int J Sci R Tech*. 2024;1(11).
138. Peña-Guerrero J, Nguewa PA, García-Sosa AT. Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. *Wiley Interdiscip Rev: Comput Mol Sci*. 2021;11(5):e1513.
139. Saraf S, De A, Tripathy BK. Effective use of computational biology and artificial intelligence in the domain of medical oncology. *Computational Intelligence for Oncology and Neurological Disorders* CRC Press; 2024:228–252.
140. Begley CG, Ashton M, Baell J, et al. Drug repurposing: Misconceptions, challenges, and opportunities for academic researchers. *Sci Transl Med*. 2021;13(612):eabd5524.
141. Qiu Y, Zhang Y, Deng Y, Liu S, Zhang W. A comprehensive review of computational methods for drug-drug interaction detection. *IEEE/ACM Trans Comput Biol Bioinform*. 2021;19(4):1968–1985.
142. Pan X, Lin X, Cao D, et al. Deep learning for drug repurposing: methods, databases, and applications. *Wiley Interdiscip Rev: Comput Mol Sci*. 2022;12(4):e1597.
143. Reang J, Yadav V, Majeed J, Sharma PC, Tonk RK, Sharma K. Informatics: tools and databases in drug discovery. *CADD and Informatics in Drug Discovery*. Singapore: Springer Nature Singapore; 2023:53–77.
144. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. *J Big Data*. 2022;9(1):102.
145. Kora P, Ooi CP, Faust O, et al. Transfer learning techniques for medical image analysis: a review. *Biocybern Biomed Eng*. 2022;42(1):79–107.
146. Cai C, Wang S, Xu Y, et al. Transfer learning for drug discovery. *J Med Chem*. 2020;63(16):8683–8694.
147. Dalkıran A, Atakan A, Rifaioğlu AS, et al. Transfer learning for drug–target interaction prediction. *Bioinformatics*. 2023;39(Supplement_1):i103–i110.
148. Iman M, Arabnia HR, Rasheed K. A review of deep transfer learning and recent advancements. *Technologies*. 2023;11(2):40.
149. Playe B, Stoven V. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. *J Cheminform*. 2020;12(1):11.
150. Dey V, Machiraju R, Ning X. Improving compound activity classification via deep transfer and representation learning. *ACS Omega*. 2022;7(11):9465–9483.
151. Yang X, Yang S, Lian X, Wuchty S, Zhang Z. Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. *Bioinformatics*. 2021;37(24):4771–4778.
152. Caniceiro AB, Orzeł U, Rosário-Ferreira N, Filipek S, Moreira IS. Leveraging artificial intelligence in GPCR activation studies: computational prediction methods as key drivers of knowledge. *Protein Supersecondary Structures: Methods and Protocols*. New York, NY: Springer US; 2024:183–220.
153. Ballester PJ, Stevens R, Haibe-Kains B, Huang RS, Aittokallio T. Artificial intelligence for drug response prediction in disease models. *Brief Bioinform*. 2022;23(1):bbab450.
154. Pachetti E, Colantonio S. A systematic review of few-shot learning in medical imaging. *Artif Intell Med*. 2024;156:102949.
155. Chegraoui H. *Machine Learning for Genomics and Imaging Data Integration Applied to Neuro-oncology*. Université Paris-Saclay; 2023.
156. Jiang J, Chen L, Ke L, et al. A review of transformers in drug discovery and beyond. *J Pharm Anal*. 2024;15(6):101081.
157. Kalusivalingam AK, Sharma A, Patel N, Singh V. Enhancing drug discovery and repurposing through transformer models and reinforcement learning algorithms. *Int J AI ML*. 2021;2(3).
158. Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G. A survey on security and privacy of federated learning. *Future Gener Comput Syst*. 2021;115:619–640.
159. Bouacida N, Mohapatra P. Vulnerabilities in federated learning. *IEEE Access*. 2021;9:63229–63249.
160. Jeyaraman N, Ramasubramanian S, Yadav S, Balaji S, Muthu S, Jeyaraman M. Regulatory challenges and frameworks for fog computing in healthcare. *Cureus*. 2024;16(8):e66779.
161. Ogunjobi TT, Ohaeri PN, Akintola OT, et al. Bioinformatics applications in chronic diseases: a comprehensive review of genomic, transcriptomics, proteomic, metabolomics, and machine learning approaches. *Medinformatics*. 2024.
162. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3(1):1–7.
163. Zheng H, Petrella JR, Doraiswamy PM, Lin G, Hao WAlzheimer's Disease Neuroimaging Initiative. Data-driven causal model discovery and personalized prediction in Alzheimer's disease. *NPJ Digit Med*. 2022;5(1):137.
164. Cremonesi F, Planat V, Kalokyri V, et al. The need for multimodal health data modeling: a practical approach for a federated-learning healthcare platform. *J Biomed Inform*. 2023;141:104338.
165. Nazeam J, Mohammed EZ, Raafat M, et al. Based on principles and insights of covid-19 epidemiology, genome sequencing, and pathogenesis: retrospective analysis of sinigrin and prolixinrx (fluphenazine) provides off-label drug candidates. *SLAS Discov: Adv Sci Drug Discov*. 2020;25(10):1123–1140.
166. Li L, Zhou Y, Ye L, Xie Z. Tracing the evolution: a comprehensive bibliometric analysis of drug interaction clinical studies. *J Clin Pharmacol*. 2024;64(12):1505–1516.
167. Agrawal S, Sarkar S, Aouedi O, et al. Federated learning for intrusion detection system: concepts, challenges and future directions. *Comput Commun*. 2022;195:346–361.
168. Rodríguez-Belenguer P, March-Vila E, Pastor M, Mangas-Sanjuan V, Soria-Olivas E. Usage of model combination in computational toxicology. *Toxicol Lett*. 2023;389:34–44.
169. Babar M, Qureshi B, Koubaa A. *Review on Federated Learning for Digital Transformation in Healthcare through Big Data Analytics*. Future Generation Computer Systems; 2024.
170. Abimannan S, El-Alfy ESM, Hussain S, et al. Towards federated learning and multi-access edge computing for air quality monitoring: literature review and assessment. *Sustainability*. 2023;15(18):13951.
171. Gangwal A, Ansari A, Ahmad I, Azad AK, Sulaiman WMAW. Current strategies to address data scarcity in artificial intelligence-based drug discovery: a comprehensive review. *Comput Biol Med*. 2024;179:108734.
172. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*. 2022;22(2):114–126.
173. Geraci J, Bhargava R, Qorri B, et al. Machine learning hypothesis-generation for patient stratification and target discovery in rare disease: our experience with Open Science in ALS. *Front Comput Neurosci*. 2024;17:1199736.
174. Hulsen T, Jamuar SS, Moody AR, et al. From big data to precision medicine. *Front Med*. 2019;6:34.
175. Tiwari PC, Pal R, Chaudhary MJ, Nath R. Artificial intelligence revolutionizing drug development: Exploring opportunities and challenges. *Drug Dev Res*. 2023;84(8):1652–1663.

176. Noor K, Lambert A. *Automation in Regulatory Reporting: AI/ML Techniques and Implementation Strategies*. EasyChair; 2024 No. 12825.
177. Singh B, Kaunert C. Embryonic machine-deep learning, smart healthcare and privacy deliberations in hospital industry: lensing confidentiality of patient's information and personal data in legal-ethical landscapes projecting futuristic dimensions. *Healthcare Industry Assessment: Analyzing Risks, Security, and Reliability*. Cham: Springer Nature Switzerland; 2024:149–170.