# Machine Learning Applications in Drug Discovery and Development

Sadettin Y. Ugurlu[a*], David McDonald[b], Shan He[c]

[a] *Department of Materials Science and Engineering, Faculty of Engineering, Akdeniz University, Antalya, Turkey*

[b] *AIA Insights Ltd, Birmingham, United Kingdom*

[c] *School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom*

s.yavuz.ugurlu@gmail.com, sxu974@alumni.bham.ac.uk

**Keywords**: *drug discovery, drug development, machine learning, artificial intelligence in drug discovery, deep learning in drug development.*

The integration of machine learning (ML) into drug discovery has significantly accelerated the identification of novel therapeutics by enabling predictive modeling across large, complex datasets. This review critically evaluates the current applications of ML in drug discovery and development, from early-stage virtual screening to clinical trial optimization. Specific tools and case studies are highlighted, including RF-Score for binding affinity prediction, LabMate.ML for chemical reaction optimization, and BioBERT for drug repurposing through natural language processing. Well-validated successes, such as AlphaFold for protein structure prediction and DeepTox for toxicity profiling, are distinguished from emerging but less validated approaches, like diffusion models and foundation models, in generative molecular design. Advanced methodologies such as geometric deep learning, explainable AI, and multi-omics integration are discussed in the context of precision medicine and multitarget drug discovery. Additionally, key challenges — such as model interpretability, data sparsity, and translational reliability — are examined to provide a realistic perspective on the role of ML in drug development. This review serves as a comprehensive and critical synthesis to guide researchers in effectively leveraging ML for impactful and scientifically grounded drug discovery.

## Introduction

Artificial Intelligence (AI) broadly refers to the field of computer science dedicated to creating systems capable of performing tasks that normally require human intelligence, such as reasoning, learning, and problem-solving [1, 2]. Machine Learning (ML), a subset of AI, focuses specifically on the development of algorithms that enable computers to learn patterns and make predictions from data without being explicitly programmed. ML techniques have become integral across many domains, including

autonomous vehicles, speech recognition, natural language processing, medical diagnostics, and drug discovery ([3, 4]).

Drug discovery and development pipelines are inherently complex and multifaceted, with a notably high failure rate often attributed to issues such as toxicity before a drug ultimately reaches the market [5, 6, 1]. This complexity underscores the urgent need for innovative strategies that can enhance efficiency and precision throughout the process. Machine learning (ML) methodologies present a powerful suite of tools capable of addressing these challenges by enabling data-driven exploration and informed decision-making based on large, high-quality datasets. By leveraging ML, the drug discovery workflow can be significantly accelerated, reducing attrition rates and improving the likelihood of clinical success.

*Scope and Significance of Machine Learning in Drug Discovery.*

Machine learning (ML) has rapidly emerged as a transformative force in drug discovery, revolutionizing traditional pipelines by enabling data-driven decision-making and predictive modeling at unprecedented scales. This review distinguishes itself by providing a comprehensive synthesis of both foundational methodologies and cutting-edge applications, spanning from early-stage virtual screening to clinical trial optimization. Emphasizing recent breakthroughs such as explainable AI, geometric deep learning, and multi-omics integration, it highlights the accelerating pace of innovation and the broadening horizons of ML-enabled therapeutics. By critically evaluating current capabilities alongside persisting challenges, this work aims to equip researchers and practitioners with an integrated understanding essential for advancing the field. The review's broad scope, coupled with a focus on emerging trends, positions it as a timely and impactful resource within the computational drug discovery landscape.

*Original Contributions and Novel Perspectives.*

While many reviews have surveyed machine learning (ML) applications in drug discovery, this work offers a distinctive contribution by integrating foundational ML principles with a critical evaluation of state-of-the-art techniques and their translational relevance. Rather than merely cataloging existing methods, the review synthesizes recent advances—such as explainable AI, geometric deep learning, and multi-omics integration—to spotlight both promising innovations and persisting bottlenecks, particularly in areas like model interpretability and multi-target drug design. Additionally, this review proposes conceptual frameworks for addressing underexplored challenges such as data scarcity, limited generalizability across biological systems, and the gap between computational predictions and experimental validation. By bridging theoretical underpinnings with applied

perspectives, the work provides a cohesive and forward-looking narrative that both complements and extends the existing literature. This integrative approach aims to equip researchers with critical insights to drive the responsible and effective application of ML in drug discovery.

**Overview and Structure of the Review**

This review is systematically structured to lead readers through the fundamental methodologies of machine learning, as well as their practical implementation across various stages of the drug discovery and development pipeline. The organization is designed to facilitate a comprehensive understanding of both theoretical concepts and applied strategies, as outlined below:

1. *Fundamental Concepts in Machine Learning and Deep Learning:* The section briefly introduces key machine learning paradigms—supervised and unsupervised learning—and highlights deep learning's role in extracting complex features from biomedical data. A clear understanding of these basics is essential for interpreting the methods and results discussed in this review.
2. *Applications of Machine Learning in Drug Discovery:*

- *Computational Virtual Screening and Molecular Property Prediction:* Explores how machine learning accelerates virtual screening and predicts molecular activity and properties.
- *In Silico Toxicity Prediction and Safety Profiling:* Describes machine learning models for early-stage toxicity assessment.
- *Data-Driven Target Identification and Validation:* Reviews machine learning strategies for discovering and validating drug targets.
- *Prediction of Ligand-Target Binding Affinities:* Examines computational approaches for accurate binding affinity estimation.
- *Machine Learning-Guided Chemical Reaction Optimization:* Presents methods to improve the efficiency and yield of synthetic reactions.
- *Druggable Binding Site Identification:* Details machine learning techniques for detecting binding pockets on proteins.
- *Computational Drug Repurposing:* Discusses the application of machine learning in identifying new therapeutic uses for existing drugs.
- *Predictive Modeling of Multi-Target Drug Candidates:* Covers approaches for identifying compounds acting on multiple targets in complex diseases.
- *Biomarker Discovery and Validation:* Summarizes machine learning methods for identifying disease and treatment response biomarkers.
- *Applications in Computational Pathology:* Highlights machine learning uses in digital pathology and image-based diagnostics.
- *Protein Structure Prediction and Stability Analysis:* Covers homology modeling and prediction of protein folding and mutation effects.
- *Pharmacokinetic and Pharmacodynamic (PK/PD) Modeling:* Explains machine learning approaches for predicting drug absorption,

distribution, metabolism, excretion, and effects.

- *Drug-Drug Interaction Prediction:* Reviews strategies to anticipate and model drug-drug interactions using machine learning.
- *Clinical Trial Simulation and Optimization:* Discusses machine learning applications in simulating and improving clinical trial design and outcomes.
- *Machine Learning in Drug Delivery System Simulation:* Outlines computational tools to simulate and optimize drug delivery mechanisms.
- *Personalized Disease Diagnosis and Treatment:* Covers machine learning applications in precision medicine and individualized therapy.
- *Cellular and Tissue-Level Computational Simulations:* Presents approaches for modeling cell and tissue dynamics using machine learning.

3. *Public Datasets and Benchmark Platforms:* Introduces widely used datasets and platforms for benchmarking ML models in drug discovery.
4. *Emerging Trends and Advanced Techniques:* Highlights new directions such as explainable AI, geometric deep learning, diffusion models, large language models, and multi-omics integration.
   - *Innovative Therapeutic Modalities Advancing Drug Discovery and Development:* Highlights cutting-edge therapeutic approaches such as KRAS inhibition, prodrug screening, PROTACs, RNA-targeting strategies, and proximity-based platforms that broaden the scope of druggable targets.
     – *Drugging KRAS:* Innovative covalent inhibitors targeting the $KRAS^{G12C}$ mutation have opened new avenues for treating previously "undruggable" oncogenes.
     – *Prodrugs Screening:* Selectively activated compounds restore or modulate difficult targets like mutant p53, enabling novel therapeutic strategies.
     – *Innovative Lead-Finding Technologies for Poorly Ligandable Targets:* Proximity-based chemoproteomic platforms identify covalent ligands and binding pockets in challenging proteins, including intrinsically disordered regions.
     – *Proximity Induction in Drug Discovery:* Molecular glues and PRO-TACs induce targeted protein degradation by harnessing endogenous ubiquitin ligases, broadening therapeutic target space.
     – *Targeting RNA as a Complement to Proteins:* Machine learning and rational design enable the identification and optimization of RNA-targeting small molecules for hard-to-treat cancers.
   - *Advanced AI Techniques Empowering the Future of Drug Discovery:* Covers transformative AI methods—including self-supervised learning, explainable AI, geometric deep learning, diffusion models, foundation models, and multi-omics integration—that drive predictive accuracy and innovation in drug development.
     – *Self-Supervised and Few-Shot Learning:* Advanced learning paradigms leverage vast

unlabeled data and limited annotations to improve molecular representation and prediction in data-scarce drug discovery tasks.

- *Explainable AI (XAI) in Drug Discovery:* Techniques that provide interpretability of complex models enhance trust and insight into AI-driven drug design decisions.
- *Geometric Deep Learning for Structure-Based Drug Discovery:* Neural networks operating on molecular graphs and 3D structures improve prediction of binding, affinity, and protein interactions.
- *Diffusion Models for De Novo Molecular Design:* Generative models based on iterative denoising produce diverse, chemically valid novel molecules tailored for specific targets.
- *Foundation Models and Large Language Models in Drug Design:* Large-scale pretrained models process multimodal chemical and biological data, enabling versatile prediction, generation, and hypothesis formulation.
- *Multi-Omics Integration Using Machine Learning:* Integrative frameworks combine diverse biological data layers to uncover disease mechanisms, predict drug response, and identify novel targets.

5. *Limitations of Machine Learning in Drug Discovery:* Highlights key challenges such as data quality and interpretability in drug discovery and development.
6. *Conclusion:* Summarizes main insights and underscores the potential of ML to transform drug discovery while noting areas for continued improvement.

**Fundamental Concepts in Machine Learning and Deep Learning**

Machine Learning (ML) is a subfield of artificial intelligence (AI) that focuses on developing algorithms and models capable of learning from data and making predictions or decisions without being explicitly programmed [7, 8], at its core, ML leverages statistical and computational methods to identify patterns within large datasets.

*Supervised learning* is the most common ML paradigm, where models are trained on labeled data (i.e., data with known input-output pairs) [9]. The algorithm learns to map inputs to outputs by minimizing error, enabling it to predict outputs for new, unseen inputs. Applications in drug discovery include activity prediction, toxicity predictions, and estimation of binding affinity.

*Unsupervised learning*, in contrast, involves modeling the structure or distribution in data without explicit labels [10, 11]. Techniques such as clustering (e.g., k-means, hierarchical clustering) and dimensionality reduction (e.g., principal component analysis, t-SNE) are used to uncover intrinsic patterns, groupings, or relationships in datasets, for example, to identify compound classes or patient subpopulations.

*Deep Learning (DL)* is a specialized subset of ML that uses neural networks with multiple layers (“deep” architectures) to automatically learn hierarchical feature representations from

raw data [12]. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [13, 14], have achieved remarkable success in image analysis, natural language processing, and molecular property prediction [15, 16, 17]. Unlike traditional ML, which often requires handcrafted feature engineering, deep learning can extract complex features directly from data, making it especially powerful for tasks such as virtual screening, activity prediction, and omics data analysis.

*Relationship between ML and DL:* All deep learning methods are a subset of ma- chine learning; however, not all machine learning methods are deep learning. While ML includes a variety of techniques such as decision trees, random forests, and support vector machines, deep learning is distinguished by its use of layered neural networks that can model highly non-linear and abstract relationships [18]. As a result, a solid grasp of these fundamental concepts—supervised learning, unsupervised learning, and the unique characteristics of deep learning—is crucial for effectively understanding the methodologies and interpreting the results commonly encountered in contemporary drug discovery research.

**Application of Machine Learning in Drug Discovery**

Machine learning (ML) addresses a series of well-defined challenges throughout the drug discovery pipeline. In each application domain, a clear problem statement defines the necessity for advanced computational solutions. This section adopts a problem-solution framework: specific challenges encountered at various drug discovery stages are concisely described, followed by a discussion of ML-based models developed to address these issues. Such an approach offers essential context, fosters diversity in methodological perspectives, and broadens the horizon for ML applications in the drug discovery and development.

*Computational Virtual Screening and Molecular Property Prediction*

A major challenge in the early stages of drug discovery is the efficient identification of promising bioactive compounds from enormous chemical libraries, which often contain millions of molecules [19, 20, 1]. Traditional experimental screening methods are costly, time-consuming, and limited by available resources, making it impractical to test more than a small fraction of possible candidates [21, 22]. Furthermore, conventional computational approaches—while faster—often rely on rigid structural rules or simplified assumptions that can overlook chemically diverse or novel actives. As a result, there is a critical need for more accurate and scalable methods that can reliably prioritize compounds with a high likelihood of biological activity while minimizing false positives and negatives [21, 19, 20]. Addressing this challenge is essential to accelerate the drug discovery process and reduce the costs associated with late-

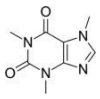
stage failures with the help of machine learning-based virtual screening approaches.

Modern drug research now depends on machine learning-based virtual screening to effectively find biologically active molecules from large chemical libraries. Two main strategies are used in this technique to find therapeutic candidates: (i) ligand-based drug design, which depends just on ligand information, and (ii) receptor-based drug design, which combines target and ligand data ([23]) (**Table 1**). By utilizing these representations (**Table 1**), a substantial amount of drug exploration has been accomplished thus far. An instance of identifying novel antibiotics was achieved through a deep neural network model, which underwent training utilizing an extensive database comprising chemical compounds and their corresponding efficacy against various bacterial strains ([24]). Another example is that the development of a generative tensorial reinforcement learning (GENTRL) model by In silico Medicine resulted in the rapid identification of new highly effective DDR1 kinase inhibitors ([25]). Also, machine learning techniques have been utilized in several applications, such as exploring potential drugs for Alzheimer's disease ([26]). The research on Alzheimer's disease showcased the capacity of machine learning-based virtual screening approaches for drug discovery.

Ligand-based drug design leverages machine learning to analyze molecular descriptors and chemical fingerprints of compounds, distinguishing between active and inactive molecules [28]. By identifying patterns from known active compounds, these models can predict the activity of newly discovered molecules. However, it is important to note that chemical similarity alone does not guarantee biological activity; for instance, molecules structurally similar to statins in DrugBank are not necessarily effective drug candidates. Ligand-based approaches are particularly useful when detailed structural information about the target receptor is unavailable or when multiple receptors with similar ligand-binding sites are of interest [28].

Conversely, receptor-based drug design leverages molecular docking simulations augmented by machine learning to predict ligand-receptor interactions with greater accuracy. In this approach, ML models are trained on structural and interaction data to evaluate the spatial and electrostatic complementarity between candidate ligands and target proteins. By learning from known binding affinities and molecular conformations, these models can efficiently prioritize compounds with high specificity and potency. This integration of docking and machine learning enables the identification of optimal binding sites and strengthens virtual screening pipelines, ultimately accelerating the selection of promising therapeutic candidates ([29]).

**Table 1**. The representation of the many techniques for representing molecular data

| Target Representation | Drug Representation | Description | Example |
|---|---|---|---|
| Sequence –Based Feature | SMILES | SMILES is Simplified Molecular Input Line Entry System it is most commonly used input to deep learning model. It represents the chemical structure used by computer can be easily learned sequence-based feature. | CC(=O) NC1 = CC = C(C = C1) O |
| | Fingerprint | It represents the molecule structure that convert into bit string i.e., presence of atoms indicates 1 and absence of atom indicates 0; it is useful method for describing the structural similarity of the molecule. | (1,1,0,0.0,1,0,1,1) |
| Structure-Based Feature | Molecular Graph | Molecular graph represents the chemical structure in terms of graph theory. It is the mapping of atoms constituting a molecule to nodes and chemical bonds to edges. |  |
| | Voxel | Voxel is the combination of volume and pixel representation of three dimensions space. It is used in the target protein because it reacts with ligand instead of the entire protein, and it is very suitable for binding prediction. | 3D target protein in cubes. |

This table categorizes commonly used molecular representations in drug discovery according to two main types: sequence-based and structure-based features. The first column defines the target representation type, while the second specifies how the drug is represented (e.g., SMILES, fingerprint, molecular graph, voxel). The third column provides a concise description of each representation technique, explaining how the molecular structure is encoded for computational use. The final column offers a corresponding example, such as a SMILES string, bit-vector fingerprint, graphical rendering, or volumetric representation. This classification aids in selecting the most appropriate input format for machine learning models based on the available molecular data and the nature of the predictive task ([27]).

Whereas receptor-based methods employ molecular docking simulations to predict binding affinities and molecular interactions, ligand-based techniques focus on assessing molecular similarity and structure-activity relationships. The combined application of these approaches enhances both the accuracy and coverage of virtual screening, thereby accelerating the identification of promising drug candidates with desired pharmacological profiles [30]. Accurate representation of drug candidates and biological targets is crucial when developing machine

learning models that integrate both ligand and receptor information [31].

In conclusion, computational virtual screening utilizing both ligand-based and receptor-based machine learning methodologies constitutes a crucial component in expediting the drug discovery process by enabling the prioritization of candidate compounds within extensive chemical libraries. Ligand-based approaches are particularly advantageous in scenarios lacking detailed receptor structural information, whereas receptor-based techniques offer mechanistic insights through modeling of molecular interactions and binding affinities. The integrated application of these complementary strategies enhances both the precision and scope of virtual screening efforts. Nonetheless, inherent limitations persist, including potential biases arising from training datasets, the complexity of accurately modeling biological environments, and the challenges associated with relying solely on chemical similarity metrics or docking scores. Therefore, rigorous validation and the integration of computational predictions with experimental data are imperative to ensure robust and translatable outcomes.

*In Silico Toxicity Prediction and Safety Profiling*

One of the most persistent and costly challenges in drug discovery is the unexpected toxicity of candidate compounds [32], which leads to a high rate of clinical trial failures and post-market drug withdrawals [33, 34]. Despite rigorous preclinical screening, many compounds that appear promising in early stages are later found to have unacceptable safety profiles, resulting in wasted resources, increased development costs, and risks to patient safety. Traditional experimental toxicity testing is time-consuming, expensive, and often limited in its ability to predict complex toxicological outcomes in humans [35, 36, 37]. This creates a critical need for more predictive and scalable methods to assess toxicity earlier in the drug development pipeline, thereby reducing late-stage attrition, improving patient safety, and restoring confidence among stakeholders.

Toxicity continues to be a significant factor contributing to the failure of medication candidates in drug development, leading to the exorbitant expenses associated with the drugs that successfully reach the market ([38]). For example, it has been observed that a significant proportion, specifically 90%, of medication candidates that have undergone clinical trials experience failure during phases I, II, or III of the trial process or throughout the process of drug licensure ([39, 40]). Despite receiving approval, some medications are subsequently pulled from the market due to their shown health concerns. This phenomenon engenders a decline in trust and assurance within the healthcare sector among patients, healthcare practitioners, investors, and regulatory bodies ([41]). Consequently, theoretical methods, such as

quantitative structure activity relationship (QSAR), offer a promising avenue to overcome these challenges by enabling earlier and more accurate toxicity prediction, thereby improving drug safety profiles and reducing late-stage attrition in the drug development pipeline ([42]).

Beyond theoretical approaches, such as QSAR, the expanding availability of comprehensive toxicity databases has significantly accelerated the adoption of diverse machine learning algorithms for predicting the toxicological profiles of small molecules. In particular, models that leverage high-throughput omics data have become increasingly instrumental in advancing toxicity prediction. Prominent examples include MolToxPred [43] and DeepTox [44], with the latter employing neural networks to deliver robust predictive performance. Recent progress in deep learning has further facilitated the incorporation of transcriptomic data, such as RNA-seq profiles, enabling highly accurate toxicity predictions, as demonstrated in **Figure 1** [45]. By integrating rich biological information, these methods offer valuable insights into potential adverse effects and support the rational design of safer therapeutic agents. Collectively, machine learning-driven toxicity prediction approaches hold substantial promise for enhancing early-stage screening and minimizing costly late-stage failures in drug discovery and development.

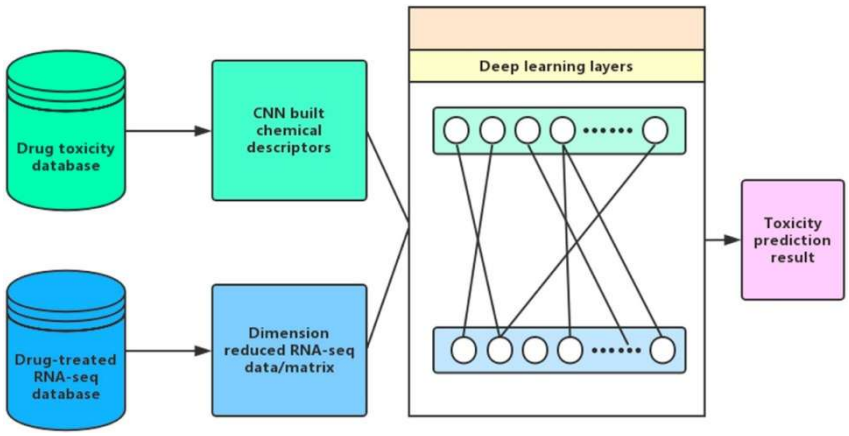



**Figure 1.** An example of ML application in RNA-seq data's toxicity prediction

RNA-seq data's toxicity prediction: a deep learning method using RNA-seq data to precisely project toxicity. This complex model uses deep learning techniques to estimate toxicity levels accurately, therefore offering important data on possible harmful effects and helping to produce safer drugs and chemicals ([45]).

Although machine learning models have demonstrated promising performance in toxicity prediction, their ability to reliably extrapolate to in vivo outcomes remains limited due to the scarcity of comprehensive in vivo data for approved drugs; most toxicity datasets originate from in vitro studies or animal models rather than human clinical trials, necessitating cautious interpretation of such predictions. Moreover, interpretability remains a significant challenge, especially for deep learning-based models, which can hinder understanding and trust in their

outputs [46, 39]. The reliability of these models in capturing true clinical toxicity is also questionable, given the differences between training data sources and complex human biology [46]. Additionally, despite the multifaceted nature of toxicity, many machine learning approaches focus on single-task predictions within isolated platforms, potentially limiting their applicability in real-world, multifactorial toxicological assessments [46].

In summary, in silico toxicity prediction has emerged as a vital and cost-effective strategy to address the persistent challenge of adverse drug effects in the drug development pipeline. While machine learning models—particularly those integrating high-throughput omics data—have shown promising accuracy in early-stage toxicity screening, significant limitations remain. The complexity of metabolic pathways, many of which remain incompletely characterized, poses a substantial challenge for model generalizability; consequently, models exhibiting excellent performance on established test datasets often experience significant drops in predictive accuracy when applied to novel compounds or pathways. This limitation underscores the importance of cautious interpretation of toxicity predictions and the necessity for continued refinement. Despite these challenges, ongoing advances in algorithm interpretability, multi-task learning, and the integration of diverse biological datasets hold considerable promise for enhancing the predictive power and clinical relevance of ML-based toxicity models. Sustained efforts in developing and validating these approaches will be crucial to fully harness their potential in improving safety assessments and reducing costly late-stage attrition in drug development.

*Data-Driven Target Identification and Validation*

Identifying effective biological targets for therapeutic intervention remains a fundamental challenge in drug discovery [47, 48]. Although advances in genomics, proteomics, and other omics technologies have yielded vast datasets, translating these data into actionable targets is not straightforward [49]. Issues such as insufficient experimental validation, biological redundancy, and unintended off-target effects often hinder the development of safe and effective therapeutics [50]. Thus, systematic prioritization and accurate prediction of druggable targets are critical to reducing attrition rates in the drug development pipeline by using traditional computational approaches.

Recent years have seen a paradigm shift in target prediction methodologies, moving from traditional computational approaches to increasingly sophisticated machine learning (ML) and deep learning (DL) models. These approaches can be categorized based on their underlying principles, data requirements, and levels of automation, as outlined below.

*i) Traditional Computational Methods*

Early computational methods for target

prediction relied primarily on established chemical and structural biology principles.

*Pharmacophore Screening:* Identifies novel targets by matching molecular features responsible for biological activity [51].

*Structure Similarity Assessment:* Utilizes chemical or structural similarity between molecules to infer potential targets [52, 53].

*Reverse Docking:* Systematically docks small molecules against a panel of protein structures to predict likely targets [54].

While these methods are valuable, their performance is highly dependent on the availability and quality of structural data, and they may not generalize well to novel compounds.

*ii) Machine Learning-Based Approaches*

Machine learning approaches for target identification and validation in drug discovery utilize diverse biological and chemical data to systematically uncover and confirm novel therapeutic targets. Supervised learning algorithms, such as Random Forests, Support Vector Machines, and Gradient Boosting Machines, have been widely applied to classify potential targets and predict their relevance based on experimental and omics datasets [55, 56]. Also, deep learning models have further enhanced target identification by automatically extracting complex hierarchical features from heterogeneous data sources. Structure-based deep learning techniques, including convolutional and graph neural networks, incorporate three-dimensional protein structures and interaction networks to more accurately predict binding sites and validate target-ligand interactions [57]. Together, these data-driven methodologies provide a comprehensive framework for accelerating the discovery and experimental validation of drug targets with high therapeutic potential.

Overall, target identification and validation have progressed significantly from relying primarily on traditional structure-based heuristics to adopting advanced machine learning and deep learning frameworks that integrate diverse data types, including multi-omics, chemical properties, and protein structures. These data-driven approaches have substantially enhanced the accuracy of target prioritization and improved the prediction of off-target effects, thereby accelerating the drug discovery and development process. Moving forward, future research should focus on developing more interpretable and generalizable models capable of integrating increasingly complex and heterogeneous biological datasets, as well as improving experimental validation pipelines to bridge the gap between computational predictions and clinical relevance. Furthermore, the incorporation of systems biology and network pharmacology concepts into machine learning frameworks holds promise for a more holistic understanding of target interactions within biological contexts. In conclusion, continued innovation in computational methodologies, combined with rigorous

experimental validation, will be essential to fully realize the potential of data-driven target discovery in delivering safer and more effective therapeutics.

*Prediction of Ligand-Target Binding Affinities*

A central challenge in drug discovery is accurately predicting how strongly a candidate molecule (ligand) will bind to its intended biological target (typically a protein) [19, 20, 58]. Binding affinity not only determines the potential efficacy of a drug but also affects its specificity and likelihood of off-target effects [59, 60, 61]. However, experimental determination of binding affinities is costly and time-consuming, requiring high-throughput screening and labor-intensive biochemical assays [62, 21]. Therefore, traditional scoring functions have been used to address these limitations, including costly and time-consuming labor-intensive biochemical assays. Unfortunately, traditional scoring functions often struggle with the complex physical and chemical factors governing molecular interactions, leading to inaccuracies in ranking or identifying optimal candidates [19, 63, 64, 65]. Therefore, more robust and predictive computational approaches, such as machine learning-based scoring approaches, are needed to efficiently prioritize compounds for further experimental validation and to reduce late-stage failures.

Machine learning-based scoring approaches have been proposed as substitutes to improve the prediction performance of conventional scoring systems. RF-Score is a prominent machine learning-based grading system ([66]). The present approach quantifies intermolecular interactions by use of atom pair counting comprising nine different heavy-atom types: C, N, O, F, P, S, Cl, Br, and I. Evaluated on the PDBbind ([67]) v2007 benchmark set, the RF-Score shows significant improvement over the existing techniques. Moreover, among the six additional parameters of the RF-Score v3 model ([68]), it shows better prediction accuracy than the original model. As described in reference ([69]), the Structural Interaction Fingerprints (SIFt) method is a machine learning technique using a format akin to fingerprints to document intermolecular interactions.

In summary, predicting ligand-target binding affinities remains a critical and challenging step in the drug discovery process. While traditional experimental methods are resource-intensive and conventional computational scoring functions often lack sufficient accuracy, machine learning-based approaches such as RF-Score and Structural Interaction Fingerprints have demonstrated promising improvements by effectively capturing complex intermolecular interactions. However, these models still face limitations, including dependence on the quality and diversity of training data, challenges in generalizing to novel targets or chemotypes, and the inherent difficulty of fully representing dynamic protein-ligand interactions. Future

research should focus on integrating multi-scale biophysical data, incorporating protein flexibility, and leveraging advances in deep learning architectures to further enhance predictive performance and interpretability. Overall, the continued development and refinement of machine learning scoring methods hold substantial potential to accelerate and improve the accuracy of drug candidate prioritization in silico.

*Machine Learning-Guided Chemical Reaction Optimization*

Optimizing chemical reactions is a critical step in drug discovery and development, directly affecting the efficiency, yield, and cost of synthesizing new pharmaceutical compounds [70, 71, 72, 73]. Traditionally, reaction optimization relies on labor-intensive trial-and-error experimentation guided by the chemist's intuition, empirical rules, or incremental improvements of known protocols [74, 75]. This process is often slow and resource-consuming and can struggle to identify optimal conditions within the vast space of possible variables (e.g., reagents, catalysts, temperature, pressure, solvents) ([76]). The complexity and combinatorial explosion of experimental parameters pose significant challenges, making it difficult to systematically explore all relevant possibilities. Therefore, there is an increasing need for more systematic, data-driven, and automated approaches, particularly machine learning (ML) models, to accelerate and improve the optimization of chemical reactions.

ML models in chemical reaction optimization can efficiently solve the restrictions of conventional approaches. Therefore, machine learning applications overcome the limitations of conventional approaches. One notable example is LabMate.ML, an automated platform for reaction optimization that iteratively improves experimental outcomes using machine learning strategies ([77]). The workflow of LabMate.ML, including its data-driven exploration and exploitation strategies for optimizing chemical reactions, is depicted in **Figure 2**.

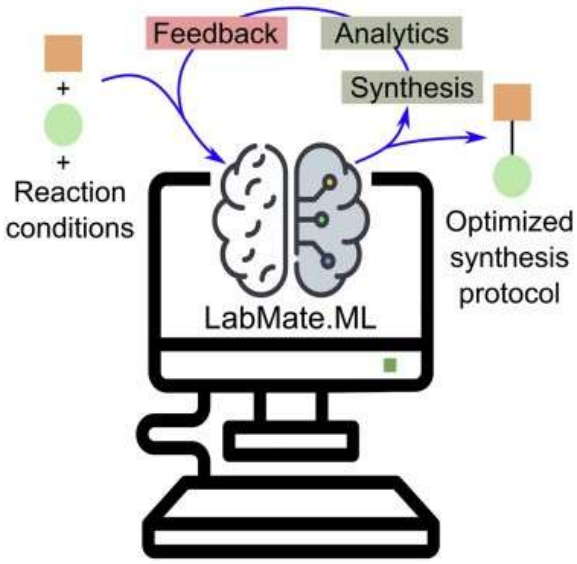



**Figure 2**. LabMate.ML Workflow for Automated Reaction Optimization

LabMate.ML simplifies the process of optimizing reaction conditions using a repetitive computational method. At first, the user inputs reactants and reaction parameters into LabMate.ML, which is shown as the left pathway in the picture. The software begins with an initial exploration phase (iterations 1-10), during which it selects starting experiments based on predictions that have the most considerable variability in output without considering the target value (measured AUC in LC-MS traces). The exploratory phase enables LabMate.ML to collect various

data points and enhance its predictive model. During iterations 11-20, LabMate.ML adopts an exploitative (greedy) strategy, prioritizing selecting conditions that maximize the target value of the reaction output. The selection approach entails selecting reactions with the least variability among the top five anticipated high-yielding reactions, given that the highest target value from the exploratory phase is four times higher than that of randomly chosen starting trials. Alternatively, if the goal value fails to exceed this threshold, LabMate.ML chooses conditions with the most significant variability among the top ten predictions that generate high results. During each iteration, LabMate.ML consistently enhances its predictive model by retraining it with fresh data points, optimizing hyperparameters (such as the number of trees, tree depth, and amount of features in RF regressors), and improving forecasts for untested circumstances that still need to be evaluated—the optimized synthesis protocol derived by LabMate.ML following repeated refinement is represented by the right pathway in the picture. This protocol enables efficient and informed experiment selection without requiring specialized gear. This method democratizes the field of synthetic chemistry by combining chemical intuition with computational automation, resulting in improved efficiency and reduced resource usage ([77]).

Beyond LabMate.ML ([77]) (Figure 2), significant advances in automated chemical design have been achieved through data-driven machine learning techniques. Graph Convolutional Neural Networks (GCNNs) have been specifically adapted for molecular representations, enabling more accurate prediction and design of chemical reactions and compounds ([78, 79, 80]). Building upon this, Kearnes et al. ([81]) introduced one-shot learning frameworks to facilitate drug discovery by rapidly generalizing from limited data. These examples demonstrate the transformative potential of ML in overcoming the constraints of traditional reaction optimization by efficiently navigating the vast chemical space and accelerating the discovery process.

In summary, machine learning-guided chemical reaction optimization offers a powerful alternative to traditional trial-and-error methods, enabling systematic, data-driven, and automated exploration of vast experimental parameter spaces. Platforms, like LabMate.ML, exemplify how iterative learning and adaptive experimentation can significantly enhance reaction yields while reducing resource consumption. Despite these advances, challenges remain, including the need for larger and more diverse high-quality datasets, the integration of domain knowledge with ML models, and the modeling of complex reaction mechanisms and dynamics. Future directions may involve combining ML with robotic automation, multi-objective optimization, and transfer learning to improve generalizability across reaction types. Overall, the integration of machine learning into chemical reaction optimization holds great promise for accelerating pharmaceutical development and expanding the frontiers of synthetic chemistry.

*Druggable Binding Site Identification*

The binding site refers to a specific localized location on a protein or biomolecule that helps other molecules, including ligands, connect and interact. Since binding sites define the drug's

efficacy, strength, and specificity in controlling biological processes, the interactions between a drug molecule and its target binding site play a crucial part in deciding the medicine's mechanism of action ([82]). Therefore, accurate identification of ligand-binding sites on protein structures is a fundamental step in drug discovery, as it directly determines the potential for effective and selective molecular interactions [19, 20, 83]. Traditional experimental approaches for mapping binding sites are time-consuming, resource-intensive, and limited by the availability of high-quality protein structures [84, 85, 19, 20, 83]. Furthermore, many proteins of pharmacological interest lack resolved structures or have flexible regions that complicate binding site detection [86, 87]. This creates a significant challenge in rapidly and reliably identifying druggable pockets, which can impede the design and development of new therapeutics. Consequently, there is a growing need for computational and data-driven binding site identification tools that can efficiently predict druggable pockets even in the absence of complete structural data, thereby accelerating drug discovery and expanding the repertoire of targetable proteins.

Computational binding site identification tools are essential for overcoming challenges in druggable pocket detection. One of the most widely used and established tools is Fpocket, which analyzes the geometric and physicochemical properties of protein surfaces to identify cavities likely to serve as ligand-binding sites. Fpocket provides rapid and reliable predictions that have been extensively validated across diverse protein targets [22]. Also, a recent approach to the identification of a binding site is the usage of structural information, with machine learning models such as dSPRINT ([88]) offering advanced predictive capabilities. dSPRINT leverages sequence and physicochemical features to identify likely ligand-binding sites on protein domains, as illustrated in **Figure 3**.

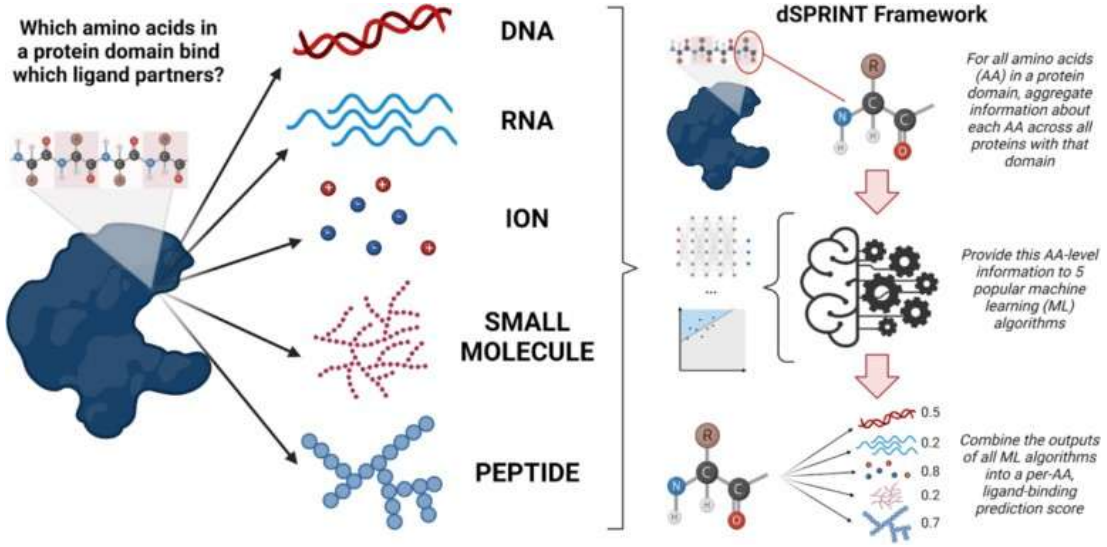



**Figure 3.** The representation of the dSPRINT (domain Sequence-based PRediction of INTeraction-sites) machine learning system for predicting protein-ligand interaction

The dSPRINT project is a machine learning-based framework designed to predict ligand-binding sites within protein domains. Utilizing a diverse set of input features—including amino acid motifs and physicochemical properties—dSPRINT employs an ensemble stacking approach to enhance predictive accuracy. Notably, the model demonstrates robust performance even in the absence of resolved crystal structures, enabling reliable identification of domain-ligand interactions in uncharacterized proteins. Cross-validation confirms the model's effectiveness in recovering known interaction sites, while its predictions also offer novel insights into protein function and molecular mechanisms. By uncovering previously unrecognized binding events, dSPRINT contributes to a deeper understanding of cellular processes and aids in the functional annotation of understudied proteins. The model architecture and its predictive outputs are publicly accessible for further exploration and evaluation ([88]).

Besides dSPRINT ([88]) (Figure 3), several structure-based approaches have been developed to find residues within protein binding sites and solve the time-consuming and expensive nature of conventional techniques ([89]). For instance, the FunFold methods use binding residues ([90, 91]) to superimpose structural templates with significant ligands onto the target model. Combining domestically developed TM-SITE and S-SITE algorithms with outside prediction tools, including COFACTOR, FINDSITE, and ConCavity, the COACH consensus approach combines One and achieves this integration by using a supervised learning technique ([92, 93]). P2Rank ([21]) is another instance of a structure-based approach applied in binding site identification. Designed especially for predicting the sites of ligand binding to protein structures, P2Rank is a powerful computer tool. P2Rank investigates protein surface physicochemical and geometric properties using machine learning (ML) approaches. This work helps to identify likely binding locations precisely. Trained on a large dataset of known protein-ligand interactions, the ML component of P2Rank gains knowledge of patterns and features separating binding sites from non-binding regions ([21]). P2Rank can generate exact forecasts even in the absence of such comparable structures. Among P2Rank's various advantages are its excellent accuracy, potency, and ability to manage several protein structures. These traits make drug discovery and development valuable. P2Rank aids in the identification of fresh binding sites, enabling the synthesis of new drugs and reusing existing compounds ([21]). At last, this accelerates the whole drug development process ([21]). The examples show that using ML for binding site predictions offers a good substitute for past methods' time-consuming and costly character.

In summary, accurate identification of druggable binding sites is a pivotal step in modern drug discovery, facilitating the design of selective and effective therapeutics. Computational tools such as Fpocket, dSPRINT, and P2Rank leverage structural, physico-chemical, and sequence-based features combined with machine learning techniques to overcome the limitations of traditional experimental methods, enabling rapid and reliable prediction of binding pockets even when structural data are

incomplete. Despite these advances, several challenges remain. Current binding site prediction tools predominantly focus on canonical orthosteric sites, whereas other important types—such as allosteric and cryptic sites—are less systematically addressed. Although individual studies have explored these alternative binding site classes, there is a noticeable gap in comprehensive methods capable of classifying and distinguishing between various binding site types. Future research should prioritize developing integrative frameworks that not only predict binding pockets but also accurately categorize them into orthosteric, allosteric, cryptic, and other relevant classes. Addressing these challenges will enhance our understanding of protein-ligand interactions and expand the druggable proteome, ultimately accelerating the discovery and development of novel therapeutics.

*Computational Drug Repurposing*

The traditional process of discovering and developing new drugs is notoriously slow, costly, and prone to high failure rates, especially during late-stage clinical trials [94, 38, 95, 96]. As a result, there is significant interest in identifying new therapeutic uses for existing drugs—a strategy known as drug repurposing or repositioning. Drug repositioning involves assigning new therapeutic indications to existing drugs that were originally developed for different diseases. A well-known example is sildenafil citrate, which was initially investigated for angina and later became widely used for treating erectile dysfunction ([97]). The concept was formalized by Ashburn and Thor in 2004, who emphasized the potential of repurposing existing drugs to reduce development time and cost while bypassing early-stage toxicity testing ([98]). Unfortunately, the major challenge in drug repurposing is efficiently and accurately predicting which existing compounds may be effective against alternative disease targets, often in the absence of extensive experimental data or balanced bioactivity datasets. Addressing this challenge requires robust machine learning models capable of identifying hidden or non-obvious relationships between drugs and diseases, despite limitations such as data sparsity, bias, and the lack of comprehensive in vitro validation.

Machine learning models have become an integral tool in drug repositioning, enabling the systematic extraction and analysis of vast biomedical data to uncover novel drug-disease associations. Recent advances leverage deep learning and natural language processing (NLP) to mine unstructured biomedical literature, extracting entities such as drugs, diseases, genes, and their relationships with high accuracy [99, 100]. For example, the MENDA system uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with attention mechanisms to predict drug-disease associations from PubMed abstracts, achieving an AUC ROC of 95.6% and successfully rediscovering known repositioning candidates like thalidomide for

leprosy [101, 100]. Similarly, the CANDO pipeline applies CNNs and long short-term memory (LSTM) networks to extract drug-cancer relationships, identifying promising repurposing candidates such as itraconazole for lung cancer [102, 100]. Moreover, during the COVID-19 pandemic, NLP models like BioBERT facilitated rapid mining of emerging literature to highlight potential therapeutics, exemplified by the identification of baricitinib as a repurposed drug candidate later validated clinically [100]. These examples illustrate how ML models, especially those integrating deep learning and NLP, offer scalable and efficient frameworks to accelerate drug repurposing by uncovering hidden and non-obvious connections that traditional methods might miss.

Despite the promising advances of machine learning models in drug repurposing, significant challenges persist. Many ML-based drug repositioning studies depend on datasets that are sparse and biased, often containing clusters of highly similar compounds such as “me too” drugs and lacking sufficient balanced examples of both active (binders) and inactive (non-binders) compounds for many targets [103, 104, 105, 106]. This data imbalance hampers the ability of ML models to generalize and accurately predict novel drug-target interactions, including off-target activities. Meaningful predictions require large, well-curated, and balanced datasets, which are typically unavailable in standard drug databases. In many cases, direct in vitro screening using deep-well plates can provide more immediate and reliable results than purely virtual ML-based methods [107, 108]. These limitations highlight the ongoing need for integrating experimental validation with computational predictions to enhance drug repurposing outcomes.

In summary, computational drug repurposing leverages advanced ML, deep learning, and NLP techniques to efficiently analyze biomedical data and uncover novel drug-disease associations, thereby accelerating the drug discovery process. Notable frameworks like MENDA, CANDO, and BioBERT-based pipelines demonstrate the potential of these technologies to identify promising repurposing candidates across therapeutic do- mains. Nevertheless, the predictive power of ML models is currently constrained by the quality and balance of available datasets, which limits generalization and accurate identification of novel targets. Future research should prioritize the development of larger, more diverse, and balanced datasets, multi-modal data integration, and interpretable ML models that can increase confidence in predictions. Furthermore, coupling ML-driven hypotheses with high-throughput experimental validation will be essential for translating computational insights into clinically effective therapies. Overall, machine learning represents a transformative approach to drug repurposing, offering opportunities to reduce costs and timelines while expanding therapeutic

options for patients worldwide.

*Predictive Modeling of Multi-Target Drug Candidates*

Traditional drug discovery has largely focused on developing compounds that selectively target a single biological molecule [109, 110, 111]. This single-target paradigm dominated drug development research over the past decades ([112]). However, selective affinity toward one target often disrupts the delicate balance of cellular metabolic networks, potentially impairing normal physiological functions and causing adverse effects [113]. Moreover, this approach may lead to limited therapeutic efficacy and unintended side effects, particularly in complex diseases characterized by multiple dysregulated pathways [114, 115]. Therefore, multi-target drugs have been developed to overcome the limitation of single-target drugs.

Chemical compounds classified as multi-target drugs combine the pharmacophores of numerous targets with different mechanisms of action into one molecule (**Figure 4**). These molecules may simultaneously interact with two or more molecular targets ([116]), which can also be called ”polypharmacology”([117]). The approach aims to target specific sets of proteins that can be modulated simultaneously by particular drugs, offering potential advantages in treating complex diseases, such as cancer ([118]). Diseases such as many cancers, neurodegenerative disorders, metabolic syndromes, and drug-induced liver injury (DILI) often involve multiple molecular targets and pathways, rendering single-target treatments insufficient [119]. However, simultaneously targeting multiple biological targets significantly increases the complexity of drug discovery, necessitating the use of advanced computational and machine learning approaches to effectively manage and analyze the resulting multidimensional data.

Advanced computational and machine learning approaches are capable of overcoming the complexity of targeting multiple biological targets by integrating diverse biological data, predicting polypharmacological profiles, and efficiently prioritizing multi-target drug candidates from large chemical libraries. The importance of multi-target drugs in managing complex diseases is well established ([120]), and recent studies by Lim et al. demonstrate the use of integrated platforms combining computational techniques and ML to identify and screen compounds with multi-target and multi-indication potential ([121]) (see **Figure 4**).

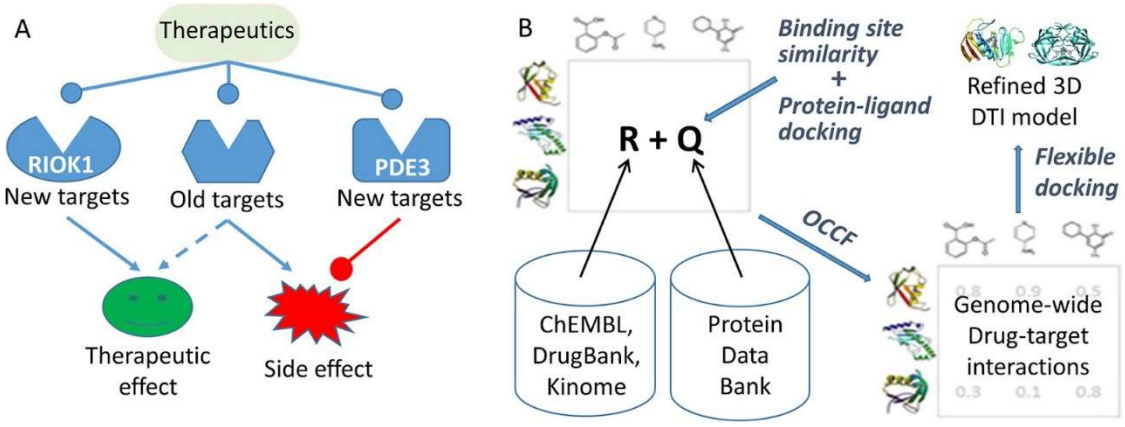



**Figure 4.** Integrated Strategy and Platform for Multi-Target Drug Screening.

(A) Illustration of a one-drug-multi-target-multi-indication strategy aimed at identifying drugs that enhance therapeutic effects while mitigating side effects, leveraging drug-off-target interactions. (B) Schema of 3D-REMAP, a robust multi-target screening platform integrating structural and chemical genomics data. The framework combines bioinformatics, chemoinformatics, protein-ligand docking, and machine-learning tools. Matrices R and Q represent observed and predicted protein-chemical interactions, respectively, and are used as inputs for the weighted imputed neighborhood-regularized One-Class Collaborative Filtering (winOCCF) algorithm to predict genome-wide drug-target interactions. Detailed methodologies are described in the Methods section ([121]).

In summary, multi-target drug design represents a promising shift away from the traditional single-target paradigm, especially for complex diseases involving multiple dysregulated pathways. By enabling simultaneous modulation of several targets, multi-target drugs hold the potential to improve therapeutic outcomes and reduce adverse effects. Machine learning models have become indispensable in this domain, facilitating the integration of heterogeneous biological datasets and the prediction of polypharmacological profiles. Nonetheless, challenges remain, including the need for more comprehensive, high-quality multi-omics data, improved interpretability of ML models, and accurate prediction of off-target effects and toxicity. Future research should emphasize enhanced data integration, the development of explainable AI methods, and robust coupling of computational predictions with experimental validation. Ultimately, the synergy of machine learning and multi-target drug discovery offers great promise to accelerate and improve treatment strategies for multifactorial diseases.

*Biomarker Discovery and Validation*

Biomarkers are measurable biological molecules or characteristics that provide critical information about physiological or pathological processes [122]. Identifying and validating reliable biomarkers is essential because they enable early disease detection, guide treatment decisions, and help monitor therapeutic effectiveness, ultimately improving patient outcomes and accelerating drug development. A major obstacle in both drug discovery and clinical practice is the timely and accurate identification of biomarkers—measurable

biological indicators that predict disease progression, drug response, or therapeutic outcomes [123, 124, 122, 125, 126]. Traditional biomarker discovery methods are often labor-intensive, require extensive experimental validation, and may fail to capture complex, multi-dimensional relationships inherent in high-throughput biological data [127, 128, 129]. Consequently, these limitations create a critical need for computational approaches, such as machine learning models, that can efficiently analyze large-scale datasets to uncover robust and clinically relevant biomarkers, thereby accelerating patient stratification, personalized therapy, and drug development.

To address this challenge, machine learning models are increasingly leveraged for biomarker discovery and drug sensitivity prediction. These computational models are particularly effective in identifying subtle, nonlinear patterns within high-dimensional datasets, enabling more accurate stratification of patients and prediction of clinical outcomes ([130, 131]). Early application of predictive ML models, utilizing preclinical or early-stage clinical trial data, can expedite biomarker validation and optimize drug development strategies ([132, 56]). Specifically, Hartman et al. ([133]) developed a pipeline that employs Binary Neural Networks (BINN) to analyze plasma proteome data from patients with acute kidney injury and COVID-19. BINNs are a class of neural networks optimized for binary (two-class) outputs, making them well-suited for distinguishing patient subphenotypes. The network is trained using protein abundance data and then interpreted with SHAP (SHapley Additive exPlanations), a model-agnostic feature importance method, to identify and prioritize significant biomarker candidates and biological pathways (see **Figure 5**) ([133]).

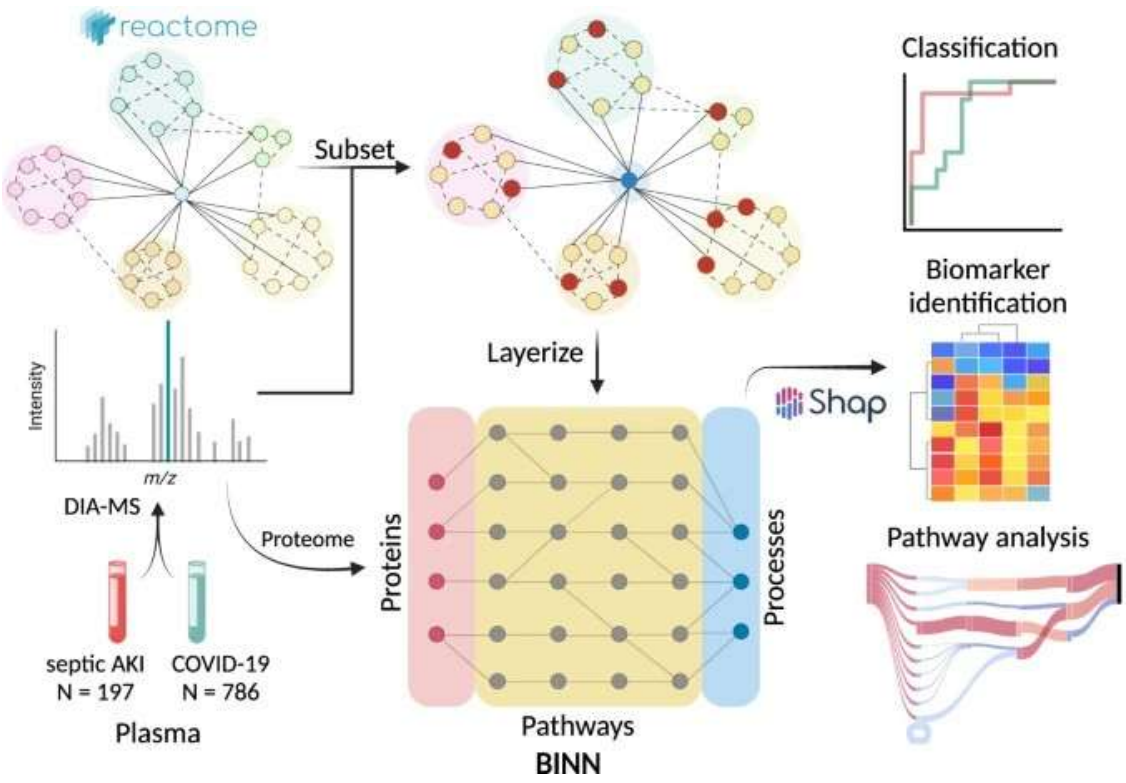



**Figure 5.** The representation of ML application in biomarker identification

The plasma proteome of patients with septic acute kidney injury (AKI) and COVID-19 was collected and examined at a different location. The data was retrieved and subjected to further analysis, which yielded datasets specific to each disease. The method commences by creating a BINN (Binary Neural Network) for each dataset. This is achieved by extracting a subset of the route database, such as Reactome, based on the proteomic content of the dataset in question. The extracted data is then organized into a sequential structure resembling a neural network. The protein amounts of each sample are utilized to train the corresponding BINNs to distinguish between two subphenotypes. Subsequently, the networks are analyzed using SHAP, which provides feature importance values that can be used to identify biomarkers and do route analysis ([133]).

As for a recent application of ML model in biomarker discovery, the IntelliGenes pipeline [134] offers a comprehensive machine learning framework for biomarker discovery and disease prediction by integrating transcriptomic, clinical, and demographic data. It combines classical statistical methods with ensemble machine learning classifiers—including Random Forest, SVM, XGBoost, k-NN, and MLP—to identify and rank disease-associated features. A notable innovation is the I-Gene score, which merges SHAP values and the Herfindahl-Hirschman Index to evaluate the predictive relevance and expression directionality of each biomarker. Applied to cardiovascular disease datasets, IntelliGenes identified 18 significant biomarkers and achieved predictive accuracies as high as 96% [134]. Its modular structure and focus on interpretability position it as a valuable tool for precision medicine and high-throughput biomarker analysis.

In conclusion, machine learning has emerged as a powerful and essential tool for biomarker discovery and validation, offering the ability to

extract meaningful patterns from high-dimensional, heterogeneous biological datasets. Techniques such as Binary Neural Networks (BINNs) have demonstrated effectiveness in distinguishing disease sub-phenotypes—such as in patients with COVID-19 and acute kidney injury—by integrating domain-specific knowledge with protein-level expression data and interpreting model outputs via SHAP. Similarly, the IntelliGenes pipeline exemplifies how combining statistical methods with ensemble learning and advanced interpretability metrics (e.g., I-Genescore) can lead to the identification of clinically relevant transcriptomic biomarkers with high predictive performance, as seen in cardiovascular disease cohorts. Despite these promising advances, several challenges persist. These include the need for larger and more diverse training datasets, robust cross-cohort validation, and better integration of multi-omics and clinical features. Additionally, model interpretability and transparency remain critical for clinical acceptance, underscoring the importance of explainable AI frameworks in biomedical applications. Future efforts should prioritize the development of scalable, user-friendly pipelines that bridge computational predictions with experimental validation, enabling more reliable translation of biomarkers into clinical diagnostics and precision therapies.

*Applications in Computational Pathology*

Traditional pathology primarily depends on manual examination of tissue samples by expert pathologists, a process that is often time-consuming, labor-intensive, and subject to inter-observer variability and human error [135, 136]. As histological and clinical datasets become increasingly large and complex, these limitations hinder consistent diagnostic accuracy and timely clinical decision-making. Therefore, there is a pressing need for computational methods, such as ML models that can systematically analyze extensive pathology data, improve reproducibility, and enhance workflow efficiency.

Computational pathology leverages ML models to extract quantitative features from histological images and other clinical data, offering more reliable and reproducible diagnostic predictions than traditional manual assessments. For example, Lee et al. demonstrated that computational analysis of non-cancerous tissue surrounding prostate tumors can predict the likelihood of cancer recurrence, providing valuable prognostic information beyond conventional pathology [137]. Similarly, studies by Lu et al. identified strong correlations between nuclear morphology and orientation with survival outcomes in oral malignancies and early-stage estrogen receptor-positive breast cancers [138, 139]. These advances illustrate how ML-driven computational pathology can reduce human error, accelerate diagnosis, and support personalized treatment decisions. Integrating computational pathology with clinical data and machine learning thus has the potential to

transform healthcare by improving diagnostic precision and enabling more targeted therapies. In conclusion, machine learning significantly enhances computational pathology by enabling the quantitative analysis of histological images and clinical data. Applications such as predicting cancer recurrence from non-cancerous tissue and leveraging nuclear morphology for prognostic assessment highlight the increasing role of ML models in improving diagnostic accuracy, minimizing human bias, and informing personalized treatment strategies. Despite its promise, computational pathology faces challenges such as the need for large, well-annotated datasets, standardization of image acquisition and processing, and model interpretability critical for clinical adoption. Additionally, integrating multi-modal data (e.g., genomics, radiology) remains complex but essential for comprehensive patient profiling. Future directions should focus on developing robust data-sharing frameworks, improving explainability of ML models to build clinician trust, and validating algorithms through prospective clinical trials. Overall, the continued evolution of computational pathology powered by machine learning stands to revolutionize diagnostics and personalized medicine, ultimately enhancing patient outcomes and healthcare efficiency.

*Protein Structure Prediction and Stability Analysis*

Protein structure prediction and stability analysis involve determining the three-dimensional arrangement of amino acids in a protein and evaluating how stable that structure is under various conditions, including mutations. These analyses are crucial because a protein's structure directly influences its function, interactions, and role in health and disease. Unfortunately, a fundamental challenge in molecular biology and drug discovery is the accurate prediction of a protein's three-dimensional (3D) structure and its stability, especially when subjected to mutations [140, 141, 142]. Since a protein's biological function is determined by its folded structure, errors in predicting this structure can impede our understanding of disease mechanisms and the development of targeted therapeutics [142]. Traditional experimental approaches such as X-ray crystallography and NMR spectroscopy are labor-intensive, time-consuming, and often infeasible for large numbers of proteins or mutated variants. Forecasting the three-dimensional configuration of a particular protein is challenging due to the vast array of possible conformations that the amino acid sequence could adopt ([143]). Therefore, there is a critical need for computational methods that can reliably forecast protein structures and assess the effects of sequence mutations on protein folding and stability in order to accelerate both basic research and drug development.

As a known common method, AlphaFold ([144]) initiates its computational exploration

method for the probable structure by utilizing a template, a pre-existing structure for similar proteins. Alternatively, AlphaFold can utilize data on the biological evolution of amino acid sequences within the protein family, which consists of proteins with similar activities and presumably possesses identical folds ([144]). The information used in AlphaFold is valuable because consistent associated evolutionary changes in pairs of amino acids suggest that these amino acids have a direct interaction despite being located far apart in the sequence ([145, 146]). Also, information of this nature can be derived from protein families' multiple sequence alignments (MSAs), obtained by analyzing the evolutionary differences in sequences across different biological species. Nevertheless, the dependence on MSAs is limiting due to the unavailability of evolutionary knowledge for all proteins ([147]).

In summary, accurate prediction of protein three-dimensional structures and their stability is essential for understanding biological function and facilitating drug discovery. Computational approaches such as AlphaFold have revolutionized this field by leveraging evolutionary information and template-based modeling to generate high-quality structural predictions, significantly reducing reliance on time-consuming experimental techniques. However, current methods face limitations, including dependence on the availability of multiple sequence alignments, challenges in modeling proteins with limited evolutionary data, and difficulties in accurately predicting the effects of mutations on dynamic stability. Future research should focus on developing algorithms capable of integrating diverse data types—such as biophysical simulations and sparse experimental inputs—and improving prediction accuracy for novel or engineered proteins. Moreover, expanding models to better capture protein dynamics and conformational flexibility will further enhance their utility in therapeutic design. Overall, continued advancements in computational protein modeling promise to accelerate biological discovery and the development of precision medicines.

*Pharmacokinetic and Pharmacodynamic (PK/PD) Modeling*

Pharmacokinetics (PK) describes how a drug is absorbed, distributed, metabolized, and excreted by the body, while pharmacodynamics (PD) focuses on the biochemical and physiological effects of the drug and its mechanism of action [148]. Together, PK/PD modeling provides critical insights into the drug's efficacy, safety, and optimal dosing regimen, playing a vital role in drug development and clinical decision-making [149, 150, 151]. Unfortunately, a major challenge in drug development is accurately predicting PK/PD profiles due to the inherent complexity and variability of biological systems. Traditional PK/PD modeling approaches often require extensive, labor-intensive experimental studies

and large datasets from animal models or clinical trials, which are costly and time-consuming [152, 153]. These limitations complicate dose optimization, adverse effect anticipation, and clinical trial design, ultimately slowing the drug development process and increasing costs. Therefore, computational methods—particularly those leveraging machine learning (ML)—are urgently needed to improve PK/PD prediction accuracy, streamline dose selection, and accelerate translation from preclinical research to clinical application.

Machine learning has shown increasing promise in PK and PK/PD modeling by enhancing predictive capabilities and expediting pharmacometric data analysis. As outlined in **Figure 6**, workflows combining traditional pharmacometric methods with ML approaches are becoming standard practice in drug development [148]. The rise of deep learning (DL) neural networks has further expanded the scope of ML applications in this field. For instance, Liu et al. (2021) utilized long short-term memory (LSTM) networks to model temporal dependencies in PK/PD data of a hypothetical drug, demonstrating improved capture of dynamic drug responses [154]. However, limitations of LSTM included challenges in extrapolating dose levels beyond the training data and handling sparse datasets. Subsequently, Liu et al. (2022) reported that artificial neural networks (ANNs) outperformed classical multinomial logistic regression in modeling drug exposure-response relationships, providing more accurate and less biased treatment effect predictions [155, 156]. These successes underscore the potential of ML to significantly improve PK/PD modeling, thereby enhancing drug discovery and development efficiency.

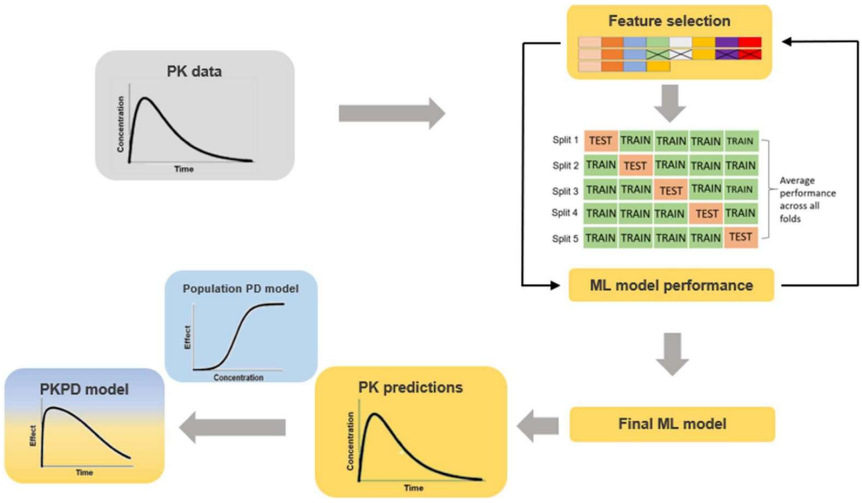



**Figure 6.** Proposed Workflow for Predicting Pharmacokinetics and Pharmacodynamics (PK/PD).

The figure illustrates a sample process that combines pharmacometrics and machine learning to provide predictions about pharmacokinetics (PK) and pharmacodynamics (PD). The blue panels represent components related to pharmacometrics, while the yellow panels indicate processes related to machine learning. The framework established in this model utilizes machine learning approaches to analyze and interpret pharmacometrics data, hence enabling PK/PD prediction. By

seamlessly integrating various methodologies, the model improves the precision and effectiveness of PK/PD prediction, assisting in optimizing drug dosage and effectiveness ([157])

In summary, pharmacokinetic and pharmacodynamic modeling are fundamental to understanding drug behavior and optimizing therapeutic regimens. Traditional methods, while foundational, face significant limitations due to biological complexity, resource-intensive data collection, and limited scalability. Machine learning approaches, particularly deep learning models such as LSTM and ANN, have demonstrated enhanced predictive accuracy and the ability to model complex temporal relationships in PK/PD data. Nevertheless, challenges remain, including the need for models that can reliably extrapolate beyond training data, manage sparse and heterogeneous datasets, and offer interpretable predictions suitable for clinical use. Future work should focus on integrating ML with mechanistic modeling, improving data quality and diversity, and developing explainable AI frameworks to foster clinical trust. Ultimately, the integration of machine learning into PK/PD modeling holds great promise to accelerate drug development and improve patient outcomes.

*Drug-Drug Interaction Prediction*

Drug-drug interactions (DDIs) occur when the concurrent administration of two or more drugs leads to alterations in their pharmacokinetics or pharmacodynamics, potentially causing reduced therapeutic efficacy, adverse drug reactions, or serious safety concerns. Predicting DDIs is therefore crucial to ensuring patient safety, guiding clinical decision-making, and optimizing drug development processes [158, 159, 160]. Traditional approaches for detecting and characterizing DDIs—such as in vitro assays, animal studies, and post-marketing surveillance—are resource-intensive, time-consuming, and insufficient to capture the growing number of potential interactions [161]. Moreover, the complexity of biological systems, involvement of drug metabolites, and variability among patient responses further complicate accurate identification of clinically relevant DDIs [162, 163]. These limitations underscore the need for scalable and data-driven computational approaches, such as machine learning models, capable of predicting DDIs efficiently, accounting for biological complexity, and supporting proactive risk assessment in both clinical and drug development.

Machine learning models offer promising solutions by integrating diverse biomedical data to systematically predict and prioritize DDIs, thereby improving drug safety assessment and informing rational combination therapy design. For example, the DeepDrug framework employs advanced deep learning architectures combining molecular and protein sequence features through residual graph convolutional and convolutional networks to predict both binary and multi-class DDIs, as well as drug-target interactions with high accuracy (see **Figure 7**) [164]. Beyond DeepDrug, various ML methods have been

proposed to forecast DDIs using publicly available databases, successfully identifying known adverse interactions and classifying interaction types [165, 166, 167]. These examples demonstrate that machine learning models hold substantial promise for DDI prediction by effectively addressing key challenges such as data sparsity, interaction complexity, and the need for scalable, high-throughput analysis.

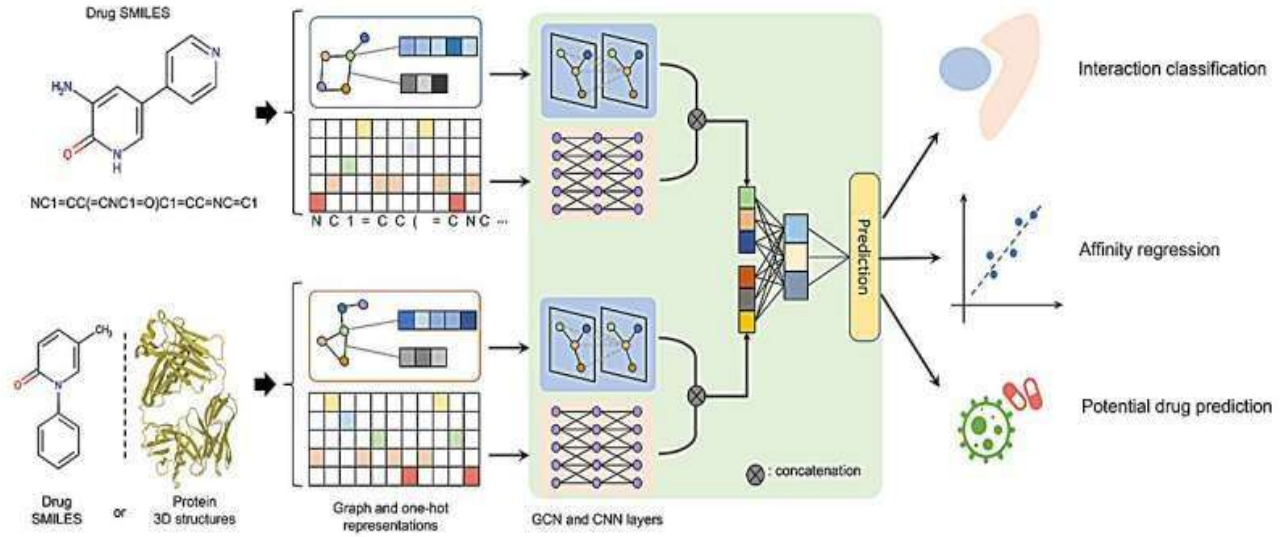



**Figure 7.** A deep learning framework is developed for predicting drug interactions and affinity.

The figure depicts a comprehensive deep-learning architecture to predict drug-drug interactions (DDI) and drug affinity regression. The DeepDrug framework, created by Frontiers Journals, combines the sequencing and structural characteristics of pharmaceuticals and proteins through residual graph convolutional networks and convolutional networks. DeepDrug performs superior to current methods in various tasks, such as binary-class drug-drug interactions (DDIs), multi-class/multi-label DDIs, binary-class drug-target interactions (DTIs) classification, and DTIs regression. The acquired structural characteristics demonstrate persistent trends in chemical attributes and drug classifications, confirming the prediction powers of DeepDrug. DeepDrug is an application that suggests prospective medications that could be effective against SARS-CoV-2. It has been validated in the relevant literature that 7 out of the top 10 ranked drugs suggested by DeepDrug are indeed promising candidates ([164]).

Despite the variety of machine learning models developed for DDI prediction, many still face persistent challenges, including data incompleteness, scarcity of reliable negative examples, and limited clinical applicability for complex decisions such as dosage optimization or alternative drug selection [160, 168, 169, 170, 171]. Also, a critical limitation for ML-based DDI prediction is the sparsity and incompleteness of drug-target affinity and metabolic data. Since many DDIs arise from drugs interacting with common targets, anti-targets, or transporters—including drug metabolites—accurate prediction requires detailed experimental profiling that is rarely available. Consequently, most ML models rely heavily on chemical structure alone or partial datasets, limiting their capacity to model complex pharmacological interactions [160]. While ML approaches have accelerated early

screening and hypothesis generation, over-reliance on computational predictions without experimental validation can lead to missed or false-positive interactions, posing risks for patient safety. Integration of ML with experimental and clinical data remains essential to maximize the reliability of DDI prediction in regulatory and clinical contexts.

In summary, predicting drug-drug interactions is a vital component of drug development and clinical pharmacology, with significant implications for patient safety and therapeutic efficacy. Traditional experimental methods, while foundational, are limited by their scalability, cost, and inability to capture the full spectrum of potential interactions. Machine learning models—especially deep learning frameworks like DeepDrug—have emerged as powerful tools to predict and classify DDIs by leveraging diverse molecular and biological data. Nevertheless, these computational methods face substantial limitations due to incomplete datasets, a lack of true-negative interaction data, and challenges in modeling metabolite-related interactions. Future research should focus on developing comprehensive, high-quality datasets including verified negative examples, improving model interpretability, and fostering integrated platforms that combine ML predictions with robust experimental validation. Enhancing the transparency and clinical applicability of ML-based DDI predictions will be critical for their adoption in regulatory decision-making and personalized medicine. Ultimately, machine learning represents a promising complementary approach to traditional methods that can help mitigate adverse drug interactions and support safer therapeutic regimens.

*Clinical Trial Simulation and Optimization*

Clinical trials are systematic investigations involving human volunteers designed to evaluate the safety, efficacy, and optimal use of novel drugs, treatments, or medical devices. Accurate clinical trial design and execution are critical for generating reliable evidence to support regulatory approval and clinical decision-making [172, 173]. However, the design and conduct of clinical trials remain among the most resource-intensive and complex stages of drug development [174, 175]. Challenges such as high financial costs, prolonged timelines, difficulties in patient recruitment, and unexpected trial terminations frequently delay or halt promising therapies. Traditional planning and simulation approaches often rely on static models and limited datasets, which restrict their ability to capture the inherent variability of patient populations, adherence, and treatment responses in real-world settings [174, 175, 176]. As the volume and complexity of clinical data increase, there is an urgent need for dynamic and sophisticated machine learning models.

Sophisticated machine learning models provide powerful tools to integrate diverse clinical, demographic, and operational data, enabling more accurate predictions of trial

outcomes, early risk identification, and data-driven optimization of trial design. Several studies have demonstrated the utility of sophisticated machine learning models in predicting clinical trial success and premature termination using both structured and unstructured data [177, 178, 179]. For example, early clinical trial replication methods primarily utilized deterministic and simple probabilistic models to simulate trial conditions and predict outcomes [180], which has been represented in **Figure 8**. The other examples are that Follett et al. applied text mining and ML methods to data from 130,000 trials, extract- ing features from study descriptions and characteristics to predict trial discontinuation [181, 182]. Similarly, Elkin et al. used document embedding, keyword analysis, and trial metadata from nearly 69,000 studies in ClinicalTrials.gov to build predictive models of trial termination [183, 184, 182]. Consequently, these examples highlight the growing role of ML in improving clinical trial simulations and reducing the time required for drug development.

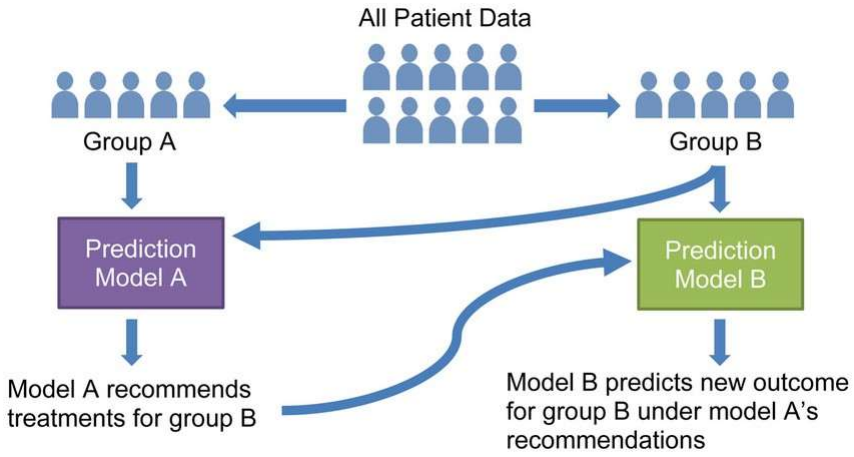



**Figure 8.** The representation of the simulation of clinical trials. Clinical trial simulation procedure. Model A suggests using in-brace corrections for patients in group B, while Model B predicts the impact of these suggestions. The predicted results were compared to the actual outcomes recorded in the charts of patients in group B ([185]).

In summary, clinical trial simulation and optimization are pivotal for enhancing the efficiency, safety, and success of drug development programs. Traditional trial design methods, though foundational, struggle with escalating costs, limited scalability, and an inability to capture real-world variability in patient adherence, recruitment, and treatment response. Machine learning-based simulation techniques offer a compelling solution by leveraging structured and unstructured data from large-scale clinical repositories to improve predictive accuracy and inform adaptive trial designs. For instance, models developed by Follett et al. and Elkin et al. demonstrated that trial discontinuation and success can be effectively predicted using metadata, natural language processing, and embedding techniques on datasets exceeding 100,000 trials. These approaches exemplify how ML can be used not

only to anticipate risk but also to proactively adjust trial parameters in silico before costly real-world implementation. Nevertheless, current models require further refinement to manage heterogeneous data sources, ensure interpretability for regulatory and clinical stakeholders, and facilitate integration with electronic trial management systems. Future research should prioritize the development of explainable AI tools, harmonized data standards, and prospective validation frameworks to ensure robust and reproducible outcomes. Overall, embedding machine learning into clinical trial design and simulation presents a transformative opportunity to accelerate the development of safe, effective, and patient-centered therapies.

*Machine Learning in Drug Delivery System Simulation*

Drug delivery system simulation involves modeling how drugs are released, distributed, and metabolized within biological systems to predict their therapeutic efficacy and safety. Accurate prediction of these processes is vital for optimizing drug formulations, ensuring targeted delivery, maximizing bioavailability, and minimizing adverse effects, thereby improving clinical outcomes [186, 187, 188]. However, despite advances in pharmaceutical research, the development and optimization of drug delivery systems remain challenging due to the complexity of biological environments and variability in patient responses [186, 189]. Traditional drug delivery simulations often depend on empirical data and simplified computational models that fail to capture the full extent of physiological variability, release kinetics, and drug distribution across heterogeneous populations. As a result, many formulations do not achieve optimal bioavailability, stability, or targeted delivery, leading to suboptimal therapeutic outcomes and increased development costs [190, 191]. These challenges underscore the need for advanced approaches—such as machine learning models—that can effectively process large, multidimensional datasets, capture com- plex nonlinear interactions, and enable predictive modeling across biological, chemical, and formulation-specific scales.

Machine learning models introduce a transformative approach to drug delivery simulations by leveraging vast databases and advanced algorithms to model complex biological systems more accurately and efficiently [192, 193]. ML-driven simulations can predict drug behavior, optimize delivery pathways, and tailor treatments to individual patient profiles, accelerating drug development while improving therapeutic effectiveness and safety [194]. The capacity of ML algorithms to analyze extensive chemical, physicochemical, and biological activity data allows for uncovering hidden relationships and nonlinear dependencies that traditional models overlook [195, 196]. Furthermore, ML models enable predictions of drug stability under various

storage conditions, solubility in diverse formulations, and release kinetics from complex delivery systems [197, 198]. By leveraging pattern recognition and trend analysis, ML-enhanced drug delivery simulations contribute to the design of more effective, safe, stable, and reliable therapeutic interventions [193, 199, 196].

In summary, drug delivery system simulation is essential for optimizing therapeutic efficacy, safety, and patient-specific treatment strategies. While traditional simulation methods provide valuable insights, they often lack the capacity to accurately model the complexity and variability of biological systems and drug formulations. Machine learning offers a powerful alternative by efficiently processing large and heterogeneous datasets, capturing nonlinear relationships, and enabling personalized predictions of drug behavior and release. However, challenges remain, including the need for high-quality, standardized datasets, the interpretability of complex ML models, and integration with mechanistic and clinical data. Future research should focus on developing hybrid modeling frame- works that combine mechanistic understanding with data-driven approaches, improving model transparency, and validating predictions through experimental and clinical studies. Overall, continued advances in machine learning hold great promise for revolutionizing drug delivery simulations and accelerating the development of safer, more effective therapeutics.

*Personalized Disease Diagnosis and Treatment*

Personalized medicine, also referred to as precision medicine, represents a paradigm shift in healthcare that aims to tailor disease identification, risk assessment, and therapeutic interventions to individual patients based on their unique genetic, molecular, environmental, and lifestyle profiles ([200, 201, 202]). Traditional approaches often rely on population averages, resulting in generalized treatments that may not account for interpatient variability in disease mechanisms or drug response [203, 204]. Recent advances in high-throughput sequencing, multi-omics technologies, and electronic health records (EHRs) have facilitated the generation of rich, multidimensional patient datasets. However, extracting actionable insights from these complex data sources remains a major challenge—one that can be effectively addressed through machine learning models capable of uncovering hidden patterns, modeling individual variability, and supporting personalized clinical decision-making.

Machine learning models are increasingly leveraged to integrate and analyze the diverse data sources to identify disease subtypes, predict disease onset, and recommend individualized treatments ([205, 206]). For example, ML models have been developed to analyze genomic and transcriptomic data for the early detection of cancer subtypes, enabling targeted screening and

intervention ([207, 208, 209]). Furthermore, ML-powered risk prediction models have been applied in cardiology, oncology, and metabolic diseases, facilitating the timely identification of high-risk patients and personalized prevention strategies ([210, 211]). Also, in the realm of therapeutic recommendations, ML-driven decision support systems utilize patient-specific molecular and clinical features to match individuals with the most effective therapies while minimizing the risk of adverse events ([212, 213]). Pharmacogenomics-guided treatment selection, immunotherapy response prediction, and adaptive dosing regimens are just a few applications where ML models are already demonstrating clinical value ([214, 215]). Despite these advances, challenges remain, including data heterogeneity, limited interpretability of complex models, and ensuring patient privacy. Nevertheless, the integration of ML with comprehensive patient data is paving the way toward highly personalized and effective disease management strategies.

In summary, personalized medicine leverages machine learning to transform disease diagnosis and treatment by integrating diverse patient-specific data, including genetic, molecular, and clinical information. This approach enables the identification of disease subtypes, early risk prediction, and tailored therapeutic recommendations, leading to improved patient outcomes and more efficient healthcare delivery. Despite these promising advances, challenges remain, such as managing data heterogeneity, enhancing the interpretability of complex ML models, and safeguarding patient privacy. Future research should focus on developing robust, explainable algorithms, improving multi-modal data integration, and establishing standardized protocols to ensure ethical and secure use of patient data. Overall, the continued convergence of machine learning with personalized healthcare holds significant potential to revolutionize disease management and usher in a new era of precision medicine.

*Cellular and Tissue-Level Computational Simulations*

Simulating cell and tissue behavior computationally is a cornerstone of systems biology and computational medicine, providing insights into complex biological processes that underpin disease progression and therapeutic response. Traditional experimental approaches, while invaluable, are often time-consuming and resource-intensive and may not capture the full spectrum of dynamic interactions present in living systems ([216, 217, 218]). Consequently, there is a growing reliance on in silico simulations that can model cellular and tissue-level phenomena under various physiological and pathological conditions.

Recent advances in mathematical modeling, agent-based simulations, and machine learning have significantly enhanced the accuracy and scalability of cell and tissue simulations ([219, 220, 221]). ML techniques, such as deep learning

and reinforcement learning, have been utilized to construct models that capture spatial and temporal dynamics of cell growth, differentiation, migration, and tissue organization, integrating data from microscopy, omics, and biophysical measurements ([222, 223]). For example, convolutional neural networks have been applied to analyze histopathological images and simulate tissue morphology, facilitating the study of tumor microenvironments and tissue regeneration ([224, 225]). Additionally, ML-augmented simulations are being used to predict the effects of genetic mutations, drug exposure, and mechanical stress on cellular behavior, supporting personalized treatment planning and biomaterial design ([226, 227]).

In clinical research, multi-scale modeling approaches that link molecular, cellular, and tissue-level data are becoming increasingly important for understanding disease mechanisms, predicting drug response, and optimizing therapeutic interventions ([228, 229]). Despite current challenges, such as computational complexity and the need for high-quality training data, the integration of machine learning with cell and tissue simulations holds immense promise for advancing precision medicine and accelerating the development of innovative therapies.

In summary, computational simulation of cell and tissue behavior has become a vital tool in drug discovery and development, enabling deeper insights into complex biological dynamics that drive disease progression and therapeutic responses. Advances in mathematical modeling, agent-based approaches, and machine learning—especially deep learning techniques—have enhanced the accuracy, scalability, and applicability of these simulations across multiple biological scales. However, challenges remain, including the high computational demands of multi-scale models, the scarcity of high-quality and comprehensive training datasets, and the difficulty of integrating heterogeneous data types. Future work should prioritize developing more efficient algorithms, improving data integration from diverse experimental modalities, and fostering standardized frameworks for validation and reproducibility. Ultimately, the continued fusion of machine learning with cell and tissue simulations promises to accelerate precision medicine efforts and drive innovative therapeutic development.

**Public Datasets and Benchmark Platforms in ML-Based Drug Discovery**

The progress of machine learning (ML) in drug discovery critically depends on the availability of large, high-quality public datasets and benchmark platforms. These resources provide standardized and curated chemical, biological, and structural data that enable the training, evaluation, and comparison of novel ML algorithms. Well-established benchmarks also facilitate reproducibility and transparent

performance assessment across different computational approaches. **Table 2** summarizes some of the most widely used datasets and benchmarking platforms in ML-driven drug discovery.

**Table 2.** Key public datasets and benchmarking platforms in ML-based drug discovery

| Dataset/Platform | Description | Typical Use/Scope |
|---|---|---|
| **PubChem [230]** | Large public repository of chemical molecules and their biological activities, maintained by NIH. | Chemical structure search, bioassay data, ML for activity prediction |
| **ChEMBL [231]** | Curated database of bioactive drug-like molecules with binding, functional, and ADMET information. | ML for bioactivity prediction, target identification, QSAR modelling |
| **BindingDB [232]** | Database of measured binding affinities for protein-ligand complexes, covering many targets | Binding affinity prediction, virtual screening, ML model benchmarking |
| **PDB (Protein Data Bank) [233]** | Central archive of 3D structural data for biological macromolecules, including proteins and nucleic acids. | Protein structure analysis, docking, ML for structural biology |
| **PDBBind [234]** | Collection of experimentally measured binding affinities and structures of protein-ligand complexes, derived from the PDB. | Training/benchmarking ML models for binding affinity/regression tasks |
| **ZINC [235]** | Free database of commercially available compounds for virtual screening. | Ligand selection, chemical library design, ML for hit identification |
| **DUDE (Directory of Useful Decoys: Enhanced) [236]** | Benchmark dataset for structure-based virtual screening, providing active compounds and property-matched decoys for multiple targets. | Docking, virtual screening, evaluation of scoring functions |
| **CASF-2016 [237]** | Comprehensive benchmarking set for scoring functions, derived from PDBBind. Contains diverse protein-ligand complexes with binding data. | Benchmarking docking and scoring functions, ML regression/classification |
| **GDB-17 [238]** | Generated database of all possible stable organic molecules up to 17 atoms of C, N, O, S, and halogens. | Chemical space exploration, generative models, de novo design |
| **DrugBank [239]** | Integrated database combining detailed drug data with comprehensive drug target information. | Drug repurposing, ML for drug-target interaction, ADMET studies |

The table summarizes major open-access datasets and benchmark platforms that have been widely adopted in the development and validation of machine learning algorithms for drug discovery and development. These resources provide chemical, biological, and structural data essential for tasks such as virtual screening, binding affinity prediction, drug-target interaction modeling, and other related applications.

Publicly available datasets (**Table 2**) continue to drive innovation in ML-based drug discovery by providing researchers with standardized, high-quality data for training,

validation, and benchmarking predictive models.

**Emerging Trends in Drug Discovery**

Recent advances in genomics, proteomics, chemoinformatics, and artificial intelligence have catalyzed a paradigm shift in drug discovery, giving rise to novel therapeutic strategies and transforming our understanding of disease biology. These emerging directions—ranging from drugging previously "undruggable" targets to leveraging generative models and multimodal learning—are critical for shaping the future of pharmaceutical innovation. Notably, many of these cutting-edge approaches remain either underexplored or only partially integrated with modern machine learning (ML) techniques. Integrating ML into these domains holds the potential to uncover latent patterns, optimize therapeutic design, and accelerate the transition from molecular insight to clinical application.

To this end, the following discussion is organized into two main sections: (i) Innovative Therapeutic Modalities Advancing Drug Discovery and Development, which explores novel chemical and biological strategies targeting previously inaccessible pathways, and (ii) Advanced AI Techniques Empowering the Future of Drug Discovery, which highlights computational and algorithmic innovations that are redefining predictive modeling and molecular design. Together, these two sections offer a comprehensive overview of how machine learning can synergistically intersect with emerging therapeutic paradigms to drive forward the next generation of drug discovery.

*Innovative Therapeutic Modalities Advancing Drug Discovery and Development*

The section presents five emerging therapeutic strategies that exemplify innovation in addressing previously inaccessible or challenging drug targets. These include: (1) Drugging KRAS, which highlights the development of covalent inhibitors against the $KRAS^{G12C}$ mutation; (2) Prodrugs Screening, focusing on selectively activated compounds that re-store functionality to mutant p53; (3) Innovative Lead-Finding Technologies for Poorly Ligandable Targets, which leverage proximity-based chemoproteomic platforms to identify covalent ligands and hidden binding pockets; (4) Proximity Induction in Drug Discovery, describing approaches such as PROTACs and molecular glues that induce targeted protein degradation; and (5) Targeting RNA as a Complement to Proteins, which explores rational and machine learning-guided strategies for designing RNA-targeting small molecules. Together, these topics underscore the expanding landscape of therapeutic modalities and their potential to reshape the boundaries of druggability

*Drugging KRAS*

KRAS is an oncogene that is commonly mutated in human cancer, with mutations occurring in around 30% of all tumors ([240]). This makes it a significant focus for drug development. However, progress in this area has been limited because there is no apparent binding pocket, and the nucleotide-binding affinity is considerable. At first, the tactics were centered on post-translational alterations. However, these strategies faced challenges due to the wide range

of substances that the enzymes involved may act upon. In 2013, Kevan Shokat made a significant discovery by devising a technique to target the mutant form of KRASG12C by covalent bonding specifically. This method involved adopting a cysteine-based tethering strategy, which led to the identification of a hidden binding pocket in the switch II domain. The study shows that KRASG12C may be selectively targeted by small compounds that specifically attach to its GDP-bound conformation, thereby offering a novel treatment strategy. Five distinct KRASG12C inhibitors are being tested in clinical trials, namely AMG-510 and MRTX849, which are now in Phase II investigations ([241, 242, 243]).

Despite the recent breakthroughs in targeting $KRAS^{G12C}$, a major limitation remains the difficulty in identifying druggable binding pockets due to the protein's intrinsically flat and dynamic surface. Traditional structure-based drug design methods, such as molecular dynamics (MD) simulations, can capture conformational flexibility and reveal cryptic sites; however, they are often computationally expensive and time-prohibitive. In this context, machine learning–accelerated molecular dynamics simulations offer a promising alternative by significantly reducing the computational burden while maintaining predictive fidelity. For instance, integrating deep learning with enhanced sampling techniques or surrogate modeling approaches can expedite the detection of transient, druggable conformations. Furthermore, machine learning models trained on density functional theory (DFT) data can emulate quantum mechanical calculations at a fraction of the computational cost, thereby enabling accurate and scalable simulations of covalent interactions such as those exploited in $KRAS^{G12C}$ inhibition. These ML-augmented frameworks can also benefit from multi-view learning paradigms that incorporate diverse inputs—such as sequence features, structural dynamics, and known ligand interactions—to enrich model generalizability and predictive power. Taken together, these innovations have the potential to overcome current limitations in targeting structurally challenging oncogenes like KRAS, accelerating the discovery of effective small-molecule inhibitors.

*Prodrugs Screening*

In cancer treatment, prodrugs like PRIMA-1 and its derivative PRIMA-1MET present a viable method for attacking ”undruggable” targets such as mutant p53. These molecules restore the active conformation of mutant p53, therefore producing sequence-specific DNA binding and cellular death. With their active metabolite methylene quinuclidine alkylating p53 and reestablishing an oxidative environment inside tumor cells, PRIMA-1 and PRIMA-1MET, most importantly, behave as prodrugs. Small compounds such as nutlins also help to solve the ”undruggable” target problem by blocking the interaction between p53 and MDM2, a negative-regulating protein that drives p53 breakdown. Nutlins offer great potential in preclinical and

clinical contexts and help to stabilize p53, therefore keeping its tumor-suppressing properties. These creative approaches draw attention to the possibilities for efficiently modifying difficult cancer targets by the use of prodrugs and tailored compounds ([244]).

The rational design of prodrugs remains a challenging task due to the need for controlled activation mechanisms, target specificity, and favorable pharmacokinetics. To address these complexities, machine learning—particularly generative modeling—offers powerful tools for expanding the prodrug chemical space and accelerating candidate discovery. Deep generative models, such as variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion-based architectures, can be trained on known prodrug datasets to generate novel molecular structures with optimized activation properties and selective reactivity. By integrating multi-objective optimization frameworks, these models can simultaneously account for desired attributes such as solubility, metabolic stability, and bioactivation pathways. Furthermore, incorporating biological context, such as tumor-specific enzymatic environments or redox conditions, into the generative process can guide the design of conditionally activated compounds tailored for specific cancer phenotypes. These ML-driven strategies not only enable the systematic exploration of vast chemical landscapes but also offer a data-driven path forward for designing next-generation prodrugs that effectively target previously inaccessible oncogenic mechanisms.

*Innovative Lead-Finding Technologies for Poorly Ligandable Targets*

The further development of proximity-based platforms will significantly influence the identification of cancer drugs. Proximity-based platforms rely on the principle that molecules or molecular entities, such as proteins or nucleic acids, in close physical proximity within a cellular or tissue environment, can interact or affect each other's function, even without direct binding. For example, building upon Shokat's research on KRASG12C using proximity-based platforms, the use of chemoproteomic methods, specifically activity-based protein profiling, has emerged as a practical approach for identifying covalent ligands and potential binding sites for proteins that are difficult to target using traditional drug-discovery methods ([245]). Also, Nomura and colleagues recently performed a cysteine-reactive covalent ligand screen to discover chemicals that can interfere with the binding of MYC to its DNA consensus sequence ([246]. On this screen, EN4 was found as a covalent ligand that specifically targets Cys-171 of MYC within a predicted intrinsically disordered area. Subsequently, this molecule was further refined to establish the first structure-activity relationships. The potential exists for ligandable pockets in numerous traditionally 'undruggable' targets to be discovered by

employing mass spectrometry techniques for covalent screening, particularly within the natural cellular environment. This approach could also enable the targeting of intrinsically disordered regions of proteins, which are commonly found in transcription factors ([240]).

Machine learning has the potential to greatly enhance proximity-based chemoproteomic approaches by improving the identification, prioritization, and functional annotation of ligandable sites in challenging protein targets. In particular, graph neural networks (GNNs) and attention-based models can be applied to structural and interaction data to predict cryptic or transient binding pockets within intrinsically disordered regions, which are often overlooked by traditional algorithms. ML models trained on mass spectrometry-derived covalent screening datasets can also be used to classify reactive amino acid residues, predict binding reactivity, and model ligand selectivity under physiological conditions. Moreover, self-supervised learning techniques—leveraging unlabeled protein-protein interaction data or proximity-labeling outputs—can generate enriched representations of protein microenvironments that capture context-specific ligandability features. Multi-modal learning frameworks integrating structural, proteomic, and transcriptomic data may further enable the prediction of functionally relevant proximity effects across diverse cellular states. Ultimately, coupling machine learning with chemoproteomic screening stands to accelerate the discovery of covalent ligands and facilitate targeted modulation of difficult proteins, particularly those residing in intrinsically disordered or transiently assembled complexes.

*Proximity Induction as a Groundbreaking New Concept in Drug Discovery*

Targeting poorly ligandable proteins such as zinc finger transcription factors, proximity induction employs small molecules to induce targeted protein degradation. This is exemplified by immunomodulatory drugs (IMiDs) and proteolysis-targeting chimeras (PRO- TACs), which facilitate selective protein breakdown by recruiting E3 ubiquitin ligases to target proteins, thereby expanding the spectrum of druggable targets [247, 248, 249]. PROTACs are heterobifunctional molecules composed of a ligand binding an E3 ligase (anchor), a ligand binding the protein of interest (warhead), and a chemical linker connecting these two. By bringing the target protein into proximity with the ubiquitination machinery, PROTACs induce its polyubiquitination and subsequent degradation by the proteasome. This modality offers distinct advantages over traditional inhibitors, including the ability to target proteins with weak binding affinities and undruggable targets, reduced drug resistance, and potential for lower toxicity [250].

A prominent example of advances in this field is the MEGA PROTAC platform, a computational pipeline designed to improve the

accuracy and efficiency of predicting PROTAC-mediated ternary complexes [250]. MEGA PROTAC leverages MEGADOCK, a rapid protein-protein docking software, combined with a sequential filtering strategy and rank aggregation to identify high-quality protein-protein complexes suitable for PROTAC docking. The method applies grid searches for translation and rotation to explore conformational space exhaustively, followed by clustering and energy-based filtering to refine candidate structures. MEGA PROTAC has demonstrated superior performance compared to state-of-the-art methods such as Bayesian optimization for ternary complex prediction (BOTCP), achieving higher DockQ scores and improved ranking efficiency across multiple benchmark cases [250]. It's publicly available code facilitates academic research and rational design of PROTACs, accelerating drug discovery against challenging targets. Overall, MEGA PROTAC exemplifies how advanced computational tools can enhance proximity-induced degradation strategies, enabling more accurate modeling of ternary complexes and facilitating the design of effective degraders against previously intractable proteins.

Despite the notable advancements introduced by MEGA PROTAC in modeling ternary complex formation, several key limitations persist that constrain its broader applicability and translational impact. One primary challenge lies in the scarcity of experimentally validated PROTAC ternary complex structures, which limits the diversity and reliability of training data for predictive modeling. This data bottleneck restricts the generalizability of structure-based tools and impairs the development of robust performance benchmarks. To address this, generative deep learning models—such as diffusion models and variational autoencoders—could be employed to augment the chemical space of PROTAC molecules, enabling in silico generation of novel linker-warhead-anchor combinations with predicted conformational feasibility. Furthermore, integrating protein structure pre- diction frameworks like AlphaFold or ESMFold may facilitate the modeling of theoretical ternary complexes in the absence of crystallographic data. Another critical gap is the lack of well-defined negative samples, which hampers the development of binary classification models for PROTAC efficacy. Establishing computational heuristics or synthetic labeling strategies to approximate true negatives could improve model calibration and performance evaluation. Overall, future research integrating machine learning with advanced structural biology holds great promise to overcome current limitations, expand the applicability of proximity-induced degradation, and accelerate rational PROTAC design.

*Targeting RNA as a Complement to Targeting Proteins*

Small-molecule treatments have attracted interest in not just targeting proteins but also in RNA regulation. For targets in difficult-to-treat

cancer, especially, this offers fresh opportunities. The promise of RNA targeting is shown by the effectiveness of PCSK9 inhibition with R-IMPP, a small molecule that binds especially to ribosomes. In clinical studies, experimental data on RNA-targeted drugs, including linezolid and ribociclin, have shown their efficacy. This implies that RNAs causing diseases can offer particular sites where medications might bind. As our knowledge of structure develops, we could be able to apply several strategies of action, including translational stalling and splicing control. Phenotypic screening techniques helped first to identify RNA-targeting drugs. However, the field is now moving towards a more intentional method known as rational design, which provides the possibility for more major clinical benefits and a better knowledge of the issue ([251, 252]).

*Application of Machine Learning Algorithms in RNA-Targeted Drug Discovery:* With the expansion of RNA structural data and RNA-ligand interaction databases, various machine learning (ML) methods are being applied to accelerate the identification and optimization of RNA-targeting compounds. For instance, support vector machines (SVM), random forests (RF), and deep learning architectures such as graph neural networks (GNNs) have been used to predict small molecule-RNA binding affinities and to classify RNA-binding sites based on sequence and structural features ([84, 253, 254, 255]). Recently, convolutional neural networks (CNNs) and transformer-based models have also been adapted for analyzing secondary and tertiary RNA structures to identify druggable pockets and to design RNA-targeted ligands ([256, 84]). Moreover, reinforcement learning has been utilized for the de novo generation of RNA-binding molecules by optimizing chemical space exploration and binding specificity. While these ML applications are still emerging, they show promising potential to complement experimental screening and rational drug design in the context of RNA-targeted therapeutics.

Despite growing interest in RNA-targeted small-molecule therapeutics, current approaches face significant limitations in accurately identifying druggable RNA structures and designing ligands with high specificity. Machine learning techniques present promising avenues to address these challenges. In particular, generative models could significantly expand the chemical space of RNA-binding compounds beyond what is accessible through phenotypic screening. These models, trained on existing RNA-ligand interaction data, could be employed to propose structurally novel scaffolds with improved binding affinity and selectivity. Moreover, given the inherent structural flexibility and context-dependence of RNA, geometric deep learning and graph-based approaches may help in identifying transient binding pockets that are often missed in conventional docking simulations. Integrating ML with transcriptomic and structural data could

also enable dynamic modeling of RNA conformational changes, aiding in the design of compounds that exploit such plasticity. Importantly, multi-view or multimodal ML frameworks—combining chemical, structural, and functional data—could provide a more holistic view of RNA-ligand interactions and facilitate rational design. As rational design becomes increasingly central to RNA-targeted drug discovery, the integration of advanced ML models stands to significantly improve prediction accuracy, reduce experimental burden, and ultimately accelerate therapeutic development against previously intractable RNA targets.

*Advanced AI Techniques Empowering the Future of Drug Discovery*

Recent advances in artificial intelligence have led to the development of powerful models capable of transforming traditional paradigms in drug discovery. This section highlights six key AI-driven approaches that are shaping the future of the field. These include: (1) Self-Supervised and Few-Shot Learning, which utilize vast amounts of unlabeled data and limited annotations to enhance molecular representation in data-scarce contexts; (2) Explainable AI (XAI), which improves transparency and interpretability of complex prediction models, fostering trust and actionable insights in drug design; (3) Geometric Deep Learning, which exploits the structural properties of molecular graphs and 3D conformations to improve binding and affinity prediction; (4) Diffusion Models for De Novo Molecular Design, which generate novel, synthetically feasible compounds through probabilistic denoising frameworks; (5) Foundation Models and Large Language Models, which integrate chemical and biological information to support hypothesis generation and multitask learning; and (6) Multi-Omics Integration, which combines heterogeneous biological datasets to elucidate complex disease mechanisms and predict therapeutic responses. Together, these techniques represent a paradigm shift toward more intelligent, interpretable, and integrative solutions in drug discovery.

*Self-Supervised and Few-Shot Learning*

Self-supervised learning (SSL) has rapidly emerged as a key paradigm in machine learning, especially for domains where labeled data is limited but vast amounts of unlabeled data are available [257, 258]. In the context of drug discovery, SSL enables the extraction of rich, informative representations from chemical structures, protein sequences, and biological networks without the need for costly manual annotations [259, 260, 261]. Typical SSL approaches involve designing pretext tasks—such as masking parts of an SMILES string or predicting missing atoms in a molecular graph—to train models to learn molecular features that generalize well to downstream tasks. Owing to its ability to leverage unlabeled data efficiently, SSL has been increasingly adopted in drug discovery and has led to several successful applications. For example, in recent years, SSL

models such as ChemBERTa and MolBERT have leveraged masked language modeling on SMILES strings to learn molecular representations, outperforming traditional descriptors in predicting chemical properties and biological activity with minimal labeled data ([262, 263, 264, 265]). Graph-based SSL methods, such as GraphCL and MolCLR, utilize contrastive learning and graph augmentation to encode the topological and chemical information of molecules, providing powerful features for tasks like virtual screening, toxicity prediction, and molecular property estimation ([266, 267, 268]). By pretraining on millions of unlabeled molecules, these models can then be fine-tuned with relatively few labeled examples for specialized tasks, accelerating the drug discovery pipeline in data-scarce scenarios.

Few-shot learning, closely related to meta-learning, aims to enable predictive modeling when only a handful of labeled examples are available—an especially relevant setting in early drug discovery or for rare disease targets. Approaches like matching networks, prototypical networks, and meta-learning frameworks have been used to transfer knowledge from well-studied targets to new proteins with limited data, greatly improving hit identification and lead optimization in low-data regimes ([269, 270, 271]). As SSL and few-shot learning methods continue to evolve, they promise to make AI-driven drug discovery more robust, data-efficient, and accessible for novel targets and therapeutic areas.

In summary, self-supervised and few-shot learning represent transformative strategies for overcoming data scarcity in drug discovery. SSL enables the extraction of generalizable molecular features from large unlabeled datasets, while few-shot learning facilitates rapid adaptation to novel tasks with minimal labeled data. These approaches have shown promising results in property prediction, virtual screening, and lead optimization. Future work should focus on developing domain-specific pretext tasks, improving transferability across molecular modalities, and integrating SSL with other learning paradigms such as reinforcement learning or multitask learning. Together, these advances hold the potential to enhance efficiency, reduce costs, and broaden the applicability of machine learning in early-stage drug discovery.

*Explainable AI (XAI) in Drug Discovery*

The adoption of complex machine learning and deep learning models in drug discovery has brought remarkable advances in prediction accuracy but has also raised concerns about interpretability and transparency [56, 272, 273]. Explainable AI (XAI) addresses this challenge by providing insights into how models reach their predictions, making AI-driven decisions more understandable and trustworthy for researchers, chemists, and regulatory authorities [274]. In drug discovery, XAI enables a deeper understanding of which molecular features or biological factors

drive model outputs, crucial for hypothesis generation, model validation, and regulatory compliance.

A wide range of XAI methods have been adopted in cheminformatics and bioinformatics. Model-intrinsic approaches, such as feature importance in tree-based models or coefficient analysis in linear models, offer straightforward interpretability. Post hoc techniques—including SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Integrated Gradients—quantify the contribution of specific input features, such as atoms, substructures, or physicochemical properties, to a model's prediction ([275, 276, 277]). In molecular graph neural networks, attention- based mechanisms and visualization tools highlight which atoms or molecular motifs are most critical for predicted binding affinity or toxicity, guiding medicinal chemists in lead optimization. Also, recent studies have shown that XAI tools can reveal non-obvious structure-activity relationships, assist in de-risking AI-designed compounds, and even identify novel functional groups that drive biological activity ([278, 279, 280]). In addition, XAI is increasingly integrated into automated drug design and generative models to provide real-time rationales for molecule suggestions. As explainable AI continues to mature, it is poised to become an essential part of the AI-enabled drug discovery workflow, enabling better collaboration between humans and algorithms, and fostering trust in AI-generated discoveries.

In conclusion, explainable AI has become a critical component in bridging the gap between high-performing machine learning models and their practical utility in drug discovery. By elucidating the contributions of molecular features and biological signals to model predictions, XAI enhances interpretability, supports regulatory transparency, and guides experimental validation. Future efforts should prioritize the development of domain-adapted XAI techniques for graph-based models, the integration of explanation metrics into model evaluation pipelines, and the use of human-in-the-loop frameworks to iteratively refine both models and hypotheses. As the complexity of AI models in drug discovery grows, explainability will remain essential to ensure their responsible and effective deployment in real-world biomedical research.

*Geometric Deep Learning for Structure-Based Drug Discovery*

Geometric deep learning (GDL) is a cutting-edge subfield that extends deep learning to non-Euclidean data, such as graphs, manifolds, and 3D molecular structures [281]. In drug discovery, GDL has revolutionized how researchers analyze and predict the behavior of molecules, proteins, and their complexes, enabling models to directly learn from the spatial and topological properties that underpin molecular recognition and binding [282, 283].

GDL models, including graph neural

networks (GNNs), equivariant neural networks, and 3D convolutional neural networks, are particularly well-suited for tasks like structure-based virtual screening, binding site prediction, and affinity estimation. For example, models like DeepDock, EquiBind, and DiffDock utilize protein and ligand 3D coordinates or molecular graphs to predict binding poses and affinities with high accuracy, outperforming traditional docking or scoring functions ([284, 285, 286, 287]). By explicitly modeling the geometric arrangement of atoms and the chemical environment, these methods can capture subtle interactions—such as hydrogen bonds, pi-stacking, and allosteric effects—that are crucial for drug efficacy.

The impact of GDL extends beyond small-molecule design. It is integral to breakthrough developments such as AlphaFold2 for protein structure prediction, allosteric site identification, and the modeling of protein-protein and protein-ligand interactions in high- throughput screening ([288]). Geometric deep learning has enabled the development of more accurate, generalizable, and physically meaningful models for structure-based drug discovery—bridging the gap between computational prediction and real-world biochemistry.

In summary, geometric deep learning has emerged as a transformative approach for structure-based drug discovery by enabling models to learn directly from the spatial and topological features of molecular and protein structures. Through architectures such as graph neural networks and equivariant models, GDL has significantly improved tasks like binding prediction, docking accuracy, and protein structure modeling. To further advance the field, future work should explore integrating GDL with generative models, enhancing interpretability, and incorporating dynamic conformational changes through molecular simulations. As these methods continue to evolve, geometric deep learning is poised to become a foundational tool in designing more precise and effective therapeutics.

*Diffusion Models for De Novo Molecular Design*

Diffusion models represent a transformative advance in generative AI, capable of creating novel, chemically valid molecules from noise via iterative denoising processes [289]. Unlike GANs or VAEs, diffusion models are particularly adept at generating diverse and high-quality samples and have recently been applied with great success to de novo molecular design and structure-based drug discovery [290, 291].

Diffusion models such as GeoDiff and DiffDock can generate new molecular structures that satisfy complex constraints, including desired physicochemical properties or specific binding site geometries ([292, 286]). These models operate by simulating a random walk from molecular noise to valid structures, learning to reverse the diffusion process and sample from the distribution of “drug-like” molecules. Conditional diffusion models can incorporate

biological information—such as protein pocket structure or activity profile—enabling the direct generation of ligands tailored for specific targets or pharmacological profiles ([293, 291, 294]).

To sum up, a key advantage of diffusion models lies in their ability to efficiently navigate vast chemical spaces, enabling the generation of structurally diverse, chemically valid, and pharmacologically relevant molecules. By leveraging iterative denoising processes and integrating conditional constraints, these models have demonstrated significant potential in de novo design, ligand generation, and binding pose prediction. Future research should focus on improving conditioning mechanisms, enhancing synthetic feasibility predictions, and integrating multi-objective optimization frameworks. As diffusion models continue to evolve, they are poised to become indispensable in the computational drug discovery toolkit, driving faster and more creative exploration of chemical space.

*Foundation Models and Large Language Models in Drug Design*

Foundation models (FMs) and large language models (LLMs) have recently transformed AI by enabling large-scale pretraining on diverse and multimodal data sources [295, 296]. In drug discovery, these models are trained on billions of chemical structures, protein sequences, biomedical literature, and even laboratory protocols, enabling them to learn generalizable representations and to perform a wide range of tasks via transfer learning or prompt-based querying [297].

Models such as ChemBERTa, MolBERT, ProteinBERT, ProtTrans, and Galactica leverage transformer architectures and self-supervised objectives to process SMILES strings, protein sequences, and other omics data ([265, 262, 298, 299, 300]). These foundation models have demonstrated state-of-the-art performance in molecular property prediction, protein-ligand interaction modeling, drug repurposing, and toxicity assessment, even when fine-tuned on limited labeled data. Beyond structured data, general-purpose LLMs (e.g., GPT-4, Med-PaLM) are being used to mine biomedical literature, automate literature reviews, and generate natural language explanations for AI predictions ([265, 262, 298, 299, 300]).

The multi-task and few-shot capabilities of foundation models enable rapid adaptation to new drug discovery tasks, such as predicting off-target effects or suggesting synthesis routes for novel compounds ([301, 302, 303]). Moreover, LLMs can facilitate human-AI collaboration by providing interactive interfaces for hypothesis generation, documentation, and experimental design. As the scale and versatility of foundation models continue to expand, they are set to become indispensable in accelerating hypothesis-driven drug discovery and in democratizing AI-powered innovation across the pharmaceutical sector. In conclusion, foundation models and large language models have emerged

as transformative tools in drug discovery by enabling large-scale, multimodal learning. Models such as ChemBERTa and MolBERT have shown strong performance in molecular property prediction using SMILES strings, while ProteinBERT and ProtTrans effectively process protein sequences for interaction modeling and function prediction. Moreover, general-purpose LLMs like GPT-4 and Med-PaLM have demonstrated their utility in literature mining, experimental documentation, and hypothesis generation. To fully realize their potential, future efforts should focus on domain-specific fine-tuning, enhanced interpretability, and seamless integration with experimental and clinical workflows. These models are poised to accelerate data-driven innovation, reduce time-to-discovery, and democratize access to AI-powered drug design.

*Multi-Omics Integration Using Machine Learning*

Multi-omics integration refers to the simultaneous analysis of multiple layers of biological information—such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics—to achieve a comprehensive understanding of disease mechanisms, drug action, and patient heterogeneity [304, 305]. In drug discovery, integrating diverse omics data presents both a tremendous opportunity and a computational challenge, as these datasets are high-dimensional, heterogeneous, and often incomplete.

Machine learning offers powerful frameworks for multi-omics data integration. Techniques such as multi-view learning, autoencoders, canonical correlation analysis (CCA), and multi-omics factor analysis (MOFA) have been successfully applied to merge disparate datasets into unified representations that reveal disease subtypes, therapeutic targets, and biomarker signatures ([304, 306, 307]). Deep learning models—including variational autoencoders and graph neural networks—can capture non-linear relationships and complex interactions across omics layers, supporting tasks such as patient stratification, drug response prediction, and mechanism-of-action elucidation.

Recent studies have demonstrated that integrating multi-omics with machine learning accelerates the identification of novel drug targets, predicts synergistic drug combinations, and supports precision medicine approaches for complex diseases like cancer ([308, 309, 310]). By modeling biological systems at multiple scales and layers, multi-omics integration is transforming drug discovery from single-gene hypothesis testing to holistic, systems-level science, enabling more accurate, interpretable, and actionable insights for therapeutic innovation.

To sum up, machine learning has become indispensable in addressing the complexity and heterogeneity inherent in multi-omics data. Approaches such as variational autoencoders,

graph neural networks, and multi-view learning have enabled the integration of genomics, transcriptomics, and proteomics to uncover disease subtypes and predict therapeutic responses. For instance, studies leveraging MOFA and deep learning have successfully stratified cancer patients and identified synergistic drug combinations. Moving forward, efforts should focus on improving data harmonization, handling missing values, and enhancing model interpretability to facilitate clinical translation. Ultimately, multi-omics integration powered by machine learning represents a vital step toward realizing personalized and systems-level drug discovery.

**Limitations of Machine Learning in Drug Discovery**

Although machine learning (ML) holds considerable promise for transforming drug discovery by enhancing predictive capabilities and accelerating workflows, several inherent challenges and limitations must be addressed to fully realize its potential.

One major challenge is the ”black-box” nature of many ML algorithms, especially deep learning models, which often lack interpretability and transparency. This opaqueness makes it difficult for researchers and regulatory authorities to understand the underlying decision-making processes, impeding trust and acceptance within the pharmaceutical industry. Developing interpretable models or integrating explainability techniques remains an active area of research crucial for the broader adoption of ML in drug development.

In addition to interpretability, practical accessibility and reproducibility represent significant hurdles. Many ML models and software tools are primarily developed and validated in Linux environments such as Ubuntu, limiting their accessibility for researchers who use Windows or macOS systems. Moreover, complex setup requirements, including dependency management and hardware compatibility, can hinder widespread practical implementation and validation in diverse laboratory settings, reducing the utility of ML approaches in real-world drug discovery pipelines.

Data quality and availability remain foundational constraints. The effectiveness of ML models depends heavily on the availability of large, diverse, and high-quality datasets. Unfortunately, such datasets are often scarce, fragmented, or biased. For example, many training datasets contain clusters of similar compounds, which can cause ML models to overfit and reduce their ability to generalize to novel chemical spaces. Moreover, the limited availability of high-quality, annotated in vivo data poses a critical barrier, as most ML models are trained on in vitro, chemical structure, or omics-derived endpoints that may not fully capture the complexity of human biology. As a result, predictions on human responses remain inherently speculative and require cautious

interpretation.

Furthermore, the transferability and generalizability of ML models are ongoing concerns. Models trained on specific datasets or experimental conditions frequently exhibit diminished performance when applied to new, unseen data or different biological contexts. Overcoming this requires collaborative efforts for standardized data collection, curation, and model validation across diverse datasets and environments.

Importantly, ML models should be viewed as tools that make relatively good predictions compared to random guessing, rather than perfect or absolute solutions. While many models have demonstrated robust validation on existing datasets, their performance on truly novel compounds or targets is less certain. The assumption that new studies will reside within the same chemical or biological space as training data must be recognized as a key caveat in interpreting ML-based predictions. Hence, integrating ML predictions with experimental validation remains essential to ensure reliable drug discovery outcomes.

**Conclusions**

Integrating machine learning (ML) into drug discovery represents a significant advancement, transforming traditional approaches and accelerating the development of new therapeutics. This review has highlighted the diverse ML methodologies currently employed in the field, including supervised learning, neural networks, and reinforcement learning, and their applications across various stages of drug discovery. By leveraging ML, researchers can enhance predictive accuracy, streamline drug design processes, and optimize clinical trials, ultimately leading to more efficient and effective drug development.

Despite the promising advancements, several challenges remain, including the need for high-quality data, model interpretability, and integration with existing workflows. Addressing these challenges will be crucial for realizing ML's full potential in drug discovery. Future developments in ML techniques and their applications promise to further improve the drug discovery pipeline, offering new opportunities for innovation and discovery.

In summary, machine learning is revolutionizing drug discovery by providing robust data analysis and prediction tools. Continued research and development in this area will undoubtedly lead to even more breakthroughs and efficiencies in the quest for novel and effective therapeutics.

A critical takeaway from this review is the necessity of nuanced interpretation: not all ML applications are equally supported by evidence. Distinguishing between validated clinical outcomes, speculative applications, and unjustified hype is essential for responsible and impactful adoption of AI in drug discovery.

**References**

[1] Ugurlu S. Computational Methods in Drug Discovery and Development. OAJ Materials and Devices

2024;8:1230-1-1230-43.

[2] Grewal P. A Critical Conceptual Analysis of Definitions of Artificial Intelligence as Applicable to Computer Engineering. IOSR Journal of Computer Engineering 2014;16(2):09-13.

[3] Zhu H. Big Data and Artificial Intelligence Modeling for Drug Discovery. Annual Review of Pharmacology and Toxicology 2020;60(1):573-589.

[4] Patel L, Shukla T, Huang X, Ussery D, Wang S. Machine Learning Methods in Drug Discovery. Molecules 2020;25(22):5277.

[5] Petrova E. Innovation in the Pharmaceutical Industry: The Process of Drug Discovery and Development. International Series in Quantitative Marketing 2013:19-81.

[6] Sinha S, Vohora D. Drug Discovery and Development. Pharmaceutical Medicine and Translational Clinical Research 2018:19-32.

[7] Sarker I. Machine Learning: Algorithms, Real-World Applications and Research Directions. 2021.

[8] Helm J, Swiergosz A, Haeberle H, Karnuta J, Schaffer J, Krebs V, Spitzer A, Ramkumar P. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. Current Reviews in Musculoskeletal Medicine 2020;13(1):69-76.

[9] Nkemdilim M, Uzoamaka P, Daniel U, Chidi M. An Overview of Supervised Machine Learning Paradigms and their Classifiers. International Journal of Advanced Engineering, Management and Science 2024;10(3):24-32.

[10] Dayan P, Sahani M, Deback G. Unsupervised learning. The MIT encyclopedia of the cognitive sciences 1999:857–859.

[11] Dike H, Zhou Y, Deveerasetty K, Wu Q. Unsupervised Learning Based On Artificial Neural Network: A Review. 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS) 2018:322-327.

[12] Wang X, Zhao Y, Pourpanah F. Recent advances in deep learning. International Journal of Machine Learning and Cybernetics 2020;11(4):747-750.

[13] Schmidt RM. Recurrent neural networks (rnns): A gentle introduction and overview. arXiv preprint 2019:1912.05911

[14] Bisong E. Recurrent Neural Networks (RNNs). Building Machine Learning and Deep Learning Models on Google Cloud Platform 2019:443-473.

[15] Mou L, Ghamisi P, Zhu X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing 2017;55(7):3639-3655.

[16] M.Tarwani K, Edem S. Survey on Recurrent Neural Network in Natural Language Processing. International Journal of Engineering Trends and Technology 2017;48(6):301-304.

[17] Walters W, Barzilay R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. Accounts of Chemical Research 2020;54(2):263-270.

[18] Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electronic Markets 2021;31(3):685-695.

[19] Ugurlu S, McDonald D, Lei H, Jones A, Li S, Tong H, Butler M, He S. Cobdock: an accurate and practical machine learning-based consensus blind docking method. Journal of Cheminformatics 2024;16(1):5.

[20] Ugurlu S. CoBdock-2: enhancing blind docking performance through hybrid feature selection combining ensemble and multimodel feature selection approaches. Journal of Computer-Aided Molecular Design 2025;39(1):48.

[21] Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of Cheminformatics 2018;10(1):1-12.

[22] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics 2009;10(1):1-11.

[23] Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, Chen G, Liu J, Jeon H, Kim M, No K, Wang G. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models.

iScience 2021;24(9):103052.

[24] Stokes J, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia N, MacNair C, French S, Carfrae L, Bloom-Ackermann Z, Tran V, Chiappino-Pepe A, Badran A, Andrews I, Chory E, Church G, Brown E, Jaakkola T, Barzilay R, Collins J. A Deep Learning Approach to Antibiotic Discovery. Cell 2020;180(4):688-702.

[25] Zhavoronkov A, Ivanenkov Y, Aliper A, Veselov M, Aladinskiy V, Aladinskaya A, Terentiev V, Polykovskiy D, Kuznetsov M, Asadulaev A, Volkov Y, Zholus A, Shayakhmetov R, Zhebrak A, Minaeva L, Zagribelnyy B, Lee L, Soll R, Madge D, Xing L, Guo T, Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology 2019;37(9):1038-1040.

[26] Carpenter K, Huang X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. Current Pharmaceutical Design 2018;24(28):3347-3358.

[27] Escalante H. Automated Machine Learning—A Brief Review at the End of the Early Years. Natural Computing Series 2021:11-28.

[28] Priya S, Tripathi G, Singh D, Jain P, Kumar A. Machine learning approaches and their applications in drug discovery and design. Chemical Biology & Drug Design 2022;100(1):136-153.

[29] Xia S, Chen E, Zhang Y. Integrated Molecular Modeling and Machine Learning for Drug Design. Journal of Chemical Theory and Computation 2023;19(21):7478-7495.

[30] Vázquez J, López M, Gibert E, Herrero E, Luque F. Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches. Molecules 2020;25(20):4723.

[31] Shiammala P, Duraimutharasan N, Vaseeharan B, Alothaim A, Al-Malki E, Snekaa B, Safi S, Singh S, Velmurugan D, Selvaraj C. Exploring the artificial intelligence and machine learning models in the context of drug design difficulties and future potential for the pharmaceutical sectors. Methods 2023;219:82-94.

[32] Segall M, Barber C. Addressing toxicity risk when designing and selecting compounds in early drug discovery. Drug Discovery Today 2014;19(5):688-693.

[33] Rao N, Powar R. Post-Marketing Drug Withdrawals: A Review. Pharmaceutical Chemistry Journal 2023;57(7):1138-1146.

[34] Kumar L. Pharmacovigilance/reporting adverse drug reactions: An approach to enhance health surveillance and extending market share by minimizing the chances of drug withdrawals. Int J Pharm Pharm Sci 2015;7(9):1–7.

[35] Krewski D, Andersen M, Tyshenko M, Krishnan K, Hartung T, Boekelheide K, Wambaugh J, Jones D, Whelan M, Thomas R, Yauk C, Barton-Maclaren T, Cote I. Toxicity testing in the 21st century: progress in the past decade and future perspectives. Archives of Toxicology 2019;94(1):1-58.

[36] Shukla S, Huang R, Austin C, Xia M. The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform. Drug Discovery Today 2010;15(23-24):997-1007.

[37] Zhang L, Zhang H, Ai H, Hu H, Li S, Zhao J, Liu H. Applications of Machine Learning Methods in Drug Toxicity Prediction. Current Topics in Medicinal Chemistry 2018;18(12):987-997.

[38] Hwang T, Carpenter D, Lauffenburger J, Wang B, Franklin J, Kesselheim A. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. JAMA Internal Medicine 2016;176(12):1826.

[39] Cavasotto C, Scardino V. Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point. ACS Omega 2022;7(51):47536-47546.

[40] Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it?. Acta Pharmaceutica Sinica B 2022;12(7):3049-3062.

[41] Andersen M, Krewski D. Toxicity Testing in the 21st Century: Bringing the Vision to Life. Toxicological Sciences 2008;107(2):324-330.

[42] Agrawal V, Khadikar P. QSAR prediction of toxicity of nitrobenzenes. Bioorganic & Medicinal

Chemistry 2001;9(11):3035-3040.

[43] Setiya A, Jani V, Sonavane U, Joshi R. MolToxPred: small molecule toxicity prediction using machine learning approach. RSC Advances 2024;14(6):4201-4220.

[44] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science 2016;3:80.

[45] Wu Y, Wang G. Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. International Journal of Molecular Sciences 2018;19(8):2358.

[46] Sharma B, Chenthamarakshan V, Dhurandhar A, Pereira S, Hendler J, Dordick J, Das P. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. Scientific Reports 2023;13(1):4908.

[47] Blundell T, Sibanda B, Montalvão R, Brewerton S, Chelliah V, Worth C, Harmer N, Davies O, Burke D. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philosophical Transactions of the Royal Society B: Biological Sciences 2006;361(1467):413-423.

[48] Schenone M, Dančík V, Wagner B, Clemons P. Target identification and mechanism of action in chemical biology and drug discovery. Nature Chemical Biology 2013;9(4):232-240.

[49] Misra B, Langefeld C, Olivier M, Cox L. Integrated omics: tools, advances and future approaches. Journal of Molecular Endocrinology 2019;62(1):R21-R45.

[50] Lopes R, Prasad M. Beyond the promise: evaluating and mitigating off-target effects in CRISPR gene editing for safer therapeutics. Frontiers in Bioengineering and Biotechnology 2024;11:1339189.

[51] Wolber G. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discovery Today 2008;13(1-2):23-29.

[52] Nettles J, Jenkins J, Bender A, Deng Z, Davies J, Glick M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. Journal of Medicinal Chemistry 2006;49(23):6802-6810.

[53] Lo Y, Senese S, Damoiseaux R, Torres J. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. ACS Chemical Biology 2016;11(8):2244-2253.

[54] Lee A, Lee K, Kim D. Using reverse docking for target identification and its applications for drug discovery. Expert Opinion on Drug Discovery 2016;11(7):707-715.

[55] Bagherian M, Sabeti E, Wang K, Sartor M, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Briefings in Bioinformatics 2020;22(1):247-269.

[56] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery 2019;18(6):463-477.

[57] Kim J, Park S, Min D, Kim W. Comprehensive Survey of Recent Drug Discovery Using Deep Learning. International Journal of Molecular Sciences 2021;22(18):9983.

[58] Surade S, Blundell T. Structural Biology and Drug Discovery of Difficult Targets: The Limits of Ligandability. Chemistry & Biology 2012;19(1):42-50.

[59] Shoshan M, Linder S. Target specificity and off-target effects as determinants of cancer drug efficacy. Expert Opinion on Drug Metabolism & Toxicology 2008;4(3):273-280.

[60] Rudmann D. On-target and Off-target-based Toxicologic Effects. Toxicologic Pathology 2012;41(2):310-314.

[61] Bender A, Scheiber J, Glick M, Davies J, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins J. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. ChemMedChem 2007;2(6):861-873.

[62] de Jong L, Uges D, Franke J, Bischoff R. Receptor–ligand binding assays: Technologies and Applications. Journal of Chromatography B 2005;829(1-

2):1-25.

[63] Tame J. Scoring functions: A view from the bench. Journal of Computer-Aided Molecular Design 1999;13(2):99-108.

[64] Li J, Fu A, Zhang L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. Interdisciplinary Sciences: Computational Life Sciences 2019;11(2):320-328.

[65] Chen Y. Beware of docking!. Trends in Pharmacological Sciences 2015;36(2):78-95.

[66] Ballester P, Mitchell J. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 2010;26(9):1169-1175.

[67] Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. Journal of Medicinal Chemistry 2004;47(12):2977-2980.

[68] Li H, Leung K, Wong M, Ballester P. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. Molecular Informatics 2015;34(2-3):115-126.

[69] Deng Z, Chuaqui C, Singh J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. Journal of Medicinal Chemistry 2003;47(2):337-344.

[70] Colombo M, Peretto I. Chemistry strategies in early drug discovery: an overview of recent trends. Drug Discovery Today 2008;13(15-16):677-684.

[71] Nicolaou K. Advancing the Drug Discovery and Development Process. Angewandte Chemie 2014;126(35):9280-9292.

[72] Campos K, Coleman P, Alvarez J, Dreher S, Garbaccio R, Terrett N, Tillyer R, Truppo M, Parmee E. The importance of synthetic chemistry in the pharmaceutical industry. Science 2019;363(6424):eaat0805.

[73] Schneider G. Automating drug discovery. Nature Reviews Drug Discovery 2017;17(2):97-113.

[74] Zhao Y, Chen G, Liu J. Polymer data challenges in the ai era: Bridging gaps for next-generation energy materials. arXiv preprint 2025:2505.13494.

[75] Saini V. Machine learning prediction of empirical polarity using SMILES encoding of organic solvents. Molecular Diversity 2022;27(5):2331-2343.

[76] Zhou Z, Li X, Zare R. Optimizing Chemical Reactions with Deep Reinforcement Learning. ACS Central Science 2017;3(12):1337-1344.

[77] Reker D, Hoyt E, Bernardes G, Rodrigues T. Adaptive Optimization of Chemical Reactions with Minimal Experimental Information. Cell Reports Physical Science 2020;1(11):100247.

[78] Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP. Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems 2015:28.

[79] Herna´ndez-Lobato JM, Requeima J, Pyzer-Knapp EO, Aspuru-Guzik A. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. In International conference on machine learning. PMLR 2017:1470–1479.

[80] Gómez-Bombarelli R, Wei J, Duvenaud D, Hernández-Lobato J, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel T, Adams R, Aspuru-Guzik A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science 2018;4(2):268-276.

[81] Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design 2016;30(8):595-608.

[82] Konc J. Binding site comparisons for target-centered drug discovery. Expert Opinion on Drug Discovery 2019;14(5):445-454.

[83] Ugurlu S, McDonald D, He S. MEF-AlloSite: an accurate and robust Multimodel Ensemble Feature

selection for the Allosteric Site identification model. Journal of Cheminformatics 2024;16(1):116.

[84] Kozlovskii I, Popov P. Computational methods for binding site prediction on macromolecules. Quarterly Reviews of Biophysics 2025;58:e12.

[85] Lexa K, Carlson H. Protein flexibility in docking and surface mapping. Quarterly Reviews of Biophysics 2012;45(3):301-343.

[86] Carlson H. Protein Flexibility is an Important Component of Structure-Based Drug Discovery. Current Pharmaceutical Design 2002;8(17):1571-1578.

[87] Sotriffer C, Klebe G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Il Farmaco 2002;57(3):243-251.

[88] Etzion-Fuchs A, Todd D, Singh M. dSPRINT: predicting DNA, RNA, ion, peptide and small molecule interaction sites within protein domains. Nucleic Acids Research 2021;49(13):e78-e78.

[89] Santana C, Izidoro S, de Melo-Minardi R, Tyzack J, Ribeiro A, Pires D, Thornton J, de A. Silveira S. GRaSP-web: a machine learning strategy to predict binding sites based on residue neighborhood graphs. Nucleic Acids Research 2022;50(W1):W392-W397.

[90] Roche D, Tetchner S, McGuffin L. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics 2011;12(1):1-20.

[91] Roche D, Buenavista M, McGuffin L. The FunFOLD2 server for the prediction of protein–ligand interactions. Nucleic Acids Research 2013;41(W1):W303-W307.

[92] Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 2013;29(20):2588-2595.

[93] Wu Q, Peng Z, Zhang Y, Yang J. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Research 2018;46(W1):W438-W442.

[94] Parasrampuria D, Benet L, Sharma A. Why Drugs Fail in Late Stages of Development: Case Study Analyses from the Last Decade and Recommendations. The AAPS Journal 2018;20(3):46.

[95] Sun A, Benet L. Late-Stage Failures of Monoclonal Antibody Drugs: A Retrospective Case Study Analysis. Pharmacology 2020;105(3-4):145-163.

[96] Tsaioun K, Bottlaender M, Mabondzo A. ADDME – Avoiding Drug Development Mistakes Early: central nervous system drug discovery perspective. BMC Neurology 2009;9(Suppl 1):S1.

[97] Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, Shi Z, Hu Z, Zhong W, Xiao G. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Research 2020;30(3):269-271.

[98] Ashburn T, Thor K. Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery 2004;3(8):673-683.

[99] R N, Khan S, kumar A, T R M, Alojail M, Sangwan S, Saraee M. Enhancing drug discovery and patient care through advanced analytics with the power of NLP and machine learning in pharmaceutical data interpretation. SLAS Technology 2025;31:100238.

[100] Thakur A, Kulkarni S, Thakur G, Khan N. Transforming drug discovery: Leveraging deep learning and nlp for accelerated drug repurposing through text mining in biomedical literature. International Journal of Intelligent Systems and Applications in Engineering 2024;21:165–172.

[101] Cai C, Guo P, Zhou Y, Zhou J, Wang Q, Zhang F, Fang J, Cheng F. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. Journal of Chemical Information and Modeling 2019;59(3):1073-1084.

[102] Madhukar N, Khade P, Huang L, Gayvert K, Galletti G, Stogniew M, Allen J, Giannakakou P, Elemento O. A Bayesian machine learning approach for drug target identification using diverse data types. Nature Communications 2019;10(1):5221.

[103] Régnier S. What is the value of ‘me-too’ drugs?.

Health Care Management Science 2013;16(4):300-313.

[104] Aronson J, Green A. Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists. British Journal of Clinical Pharmacology 2020;86(11):2114-2122.

[105] Napolitano F, Zhao Y, Moreira V, Tagliaferri R, Kere J, D'Amato M, Greco D. Drug repositioning: a machine-learning approach through data integration. Journal of Cheminformatics 2013;5(1):30.

[106] Koromina M, Pandi M, Patrinos G. Rethinking Drug Repositioning and Development with Artificial Intelligence, Machine Learning, and Omics. OMICS: A Journal of Integrative Biology 2019;23(11):539-548.

[107] Peteani G, Huynh M, Gerebtzoff G, Rodríguez-Pérez R. Application of machine learning models for property prediction to targeted protein degraders. Nature Communications 2024;15(1):5764.

[108] Khuat TT, Bassett R, Otte E, Grevis-James A, Gabrys B. Applications of machine learning in biopharmaceutical process development and manufacturing: Current trends, challenges, and opportunities. arXiv preprint 2023:2310.09991.

[109] Gupta N, Pandya P, Verma S. Computational Predictions for Multi-Target Drug Design. Methods in Pharmacology and Toxicology 2018:27-50.

[110] Feldmann C, Yonchev D, Bajorath J. Analysis of Biological Screening Compounds with Single- or Multi-Target Activity via Diagnostic Machine Learning. Biomolecules 2020;10(12):1605.

[111] Kleandrova V, DS Cordeiro M, Speck-Planche A. Current In Silico Methods for Multi-Target Drug Discovery in Early Anticancer Research: The Rise of the Perturbation-Theory Machine Learning Approach. Future Medicinal Chemistry 2023;15(18):1647-1650.

[112] Morphy R, Kay C, Rankovic Z. From magic bullets to designed multiple ligands. Drug Discovery Today 2004;9(15):641-651.

[113] Rothman R, Baumann M, Savage J, Rauser L, McBride A, Hufeisen S, Roth B. Evidence for Possible Involvement of 5-HT 2B Receptors in the Cardiac Valvulopathy Associated With Fenfluramine and Other Serotonergic Medications. Circulation 2000;102(23):2836-2841.

[114] Löscher W. Single-Target Versus Multi-Target Drugs Versus Combinations of Drugs With Multiple Targets: Preclinical and Clinical Evidence for the Treatment or Prevention of Epilepsy. Frontiers in Pharmacology 2021;12:730257.

[115] Makhoba X, Viegas Jr. C, Mosa R, Viegas F, Pooe O. Potential Impact of the Multi-Target Drug Approach in the Treatment of Some Complex Diseases. Drug Design, Development and Therapy 2020;Volume 14:3235-3249.

[116] Morphy R, Rankovic Z. Multi-target Drugs. The Practice of Medicinal Chemistry 2008:549-571.

[117] Lavecchia A, Cerchia C. In silico methods to address polypharmacology: current status, applications and future perspectives. Drug Discovery Today 2016;21(2):288-298.

[118] Anighoro A, Bajorath J, Rastelli G. Polypharmacology: Challenges and Opportunities in Drug Discovery. Journal of Medicinal Chemistry 2014;57(19):7874-7887.

[119] Real M, Barnhill M, Higley C, Rosenberg J, Lewis J. Drug-Induced Liver Injury: Highlights of the Recent Literature. Drug Safety 2018;42(3):365-387.

[120] Hopkins A. Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology 2008;4(11):682-690.

[121] Lim H, He D, Qiu Y, Krawczuk P, Sun X, Xie L. Rational discovery of dual-indication multi-target PDE/Kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology. PLOS Computational Biology 2019;15(6):e1006619.

[122] Tan D, Thomas G, Garrett M, Banerji U, de Bono J, Kaye S, Workman P. Biomarker-Driven Early Clinical Trials in Oncology. The Cancer Journal 2009;15(5):406-420.

[123] Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nature Reviews Drug Discovery 2003;2(7):566-580.

[124] Frangogiannis N. Biomarkers: hopes and challenges in the path from discovery to clinical practice. Translational Research 2012;159(4):197-204.

[125] Aydin B, Arga K, Karadag A. Omics-Driven Biomarkers of Psoriasis: Recent Insights, Current Challenges, and Future Prospects. Clinical, Cosmetic and Investigational Dermatology 2020;Volume 13:611-625.

[126] Kori M, Gov E, Arga K, Sinha R. Biomarkers From Discovery to Clinical Application: In Silico Pre-Clinical Validation Approach in the Face of Lung Cancer. Biomarker Insights 2024;19:11772719241287400.

[127] Nakayasu ES, Gritsenko M, Piehowski PD, Gao Y, Orton DJ, Schepmoes AA, Fillmore TL, Frohnert BI, Rewers M, Krischer JP. Tutorial: best practices and considerations for mass- spectrometry-based protein biomarker discovery and validation. Nature Protocols 2021;16(8):3737–3760.

[128] Clark A, Lillard J. A Comprehensive Review of Bioinformatics Tools for Genomic Biomarker Discovery Driving Precision Oncology. Genes 2024;15(8):1036.

[129] Nagana Gowda G, Raftery D. Biomarker Discovery and Translation in Metabolomics. Current Metabolomics 2013;1(3):227-240.

[130] Li B, Shin H, Gulbekyan G, Pustovalova O, Nikolsky Y, Hope A, Bessarabova M, Schu M, Kolpakova-Hart E, Merberg D, Dorner A, Trepicchio W. 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.

[131] Kraus V. Biomarkers as drug development tools: discovery, validation, qualification and use. Nature Reviews Rheumatology 2018;14(6):354-362.

[132] Weatherall J, Khan F, Patel M, Dearden R, Shameer K, Dennis G, Feldberg G, White T, Khosla S. Clinical trials, real-world evidence, and digital medicine. The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry 2021:191-215.

[133] Hartman E, Scott A, Karlsson C, Mohanty T, Vaara S, Linder A, Malmström L, Malmström J. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. Nature Communications 2023;14(1):5359.

[134] DeGroat W, Mendhe D, Bhusari A, Abdelhalim H, Zeeshan S, Ahmed Z. IntelliGenes: a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. Bioinformatics 2023;39(12):btad755.

[135] Fuchs T, Buhmann J. Computational pathology: Challenges and promises for tissue analysis. Computerized Medical Imaging and Graphics 2011;35(7-8):515-530.

[136] Lyzogub M. Overview of clinical validation processes for artificial intelli- gence applications in pathology. PhD thesis, Vilniaus universitetas 2024.

[137] Lee G, Veltri R, Zhu G, Ali S, Epstein J, Madabhushi A. Nuclear Shape and Architecture in Benign Fields Predict Biochemical Recurrence in Prostate Cancer Patients Following Radical Prostatectomy: Preliminary Findings. European Urology Focus 2017;3(4-5):457-466.

[138] Lu C, Lewis J, Dupont W, Plummer W, Janowczyk A, Madabhushi A. An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. Modern Pathology 2017;30(12):1655-1665.

[139] Lu C, Romo-Bucheli D, Wang X, Janowczyk A, Ganesan S, Gilmore H, Rimm D, Madabhushi A. Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers. Laboratory Investigation 2018;98(11):1438-1448.

[140] Schmidt T, Bergner A, Schwede T. Modelling three-dimensional protein structures for applications in drug design. Drug Discovery Today 2014;19(7):890-897.

[141] Schauperl M, Denny R. AI-Based Protein Structure Prediction in Drug Discovery: Impacts and Challenges. Journal of Chemical Information and Modeling 2022;62(13):3142-3156.

[142] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular

Cell Biology 2019;20(11):681-697.

[143] Axen S, Huang X, Cáceres E, Gendelev L, Roth B, Keiser M. A Simple Representation of Three-Dimensional Molecular Structure. Journal of Medicinal Chemistry 2017;60(17):7393-7409.

[144] David A, Islam S, Tankhilevich E, Sternberg M. The AlphaFold Database of Protein Structures: A Biologist's Guide. Journal of Molecular Biology 2022;434(2):167336.

[145] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D, Sander C, Zecchina R, Onuchic J, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 2011;108(49):E1293–E1301.

[146] Marks D, Colwell L, Sheridan R, Hopf T, Pagnani A, Zecchina R, Sander C. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS ONE 2011;6(12):e28766.

[147] Nourmohammad A, Pun M, Visani G. Machine-Learning Model Reveals Protein-Folding Physics. Physics 2022;15:183.

[148] Meibohm B, Derendorf H. Basic concepts of pharmacokinetic/pharmacodynamic (pk/pd) modelling. International journal of clinical pharmacology and therapeutics 1997;35(10):401–413.

[149] Zou H, Banerjee P, Leung S, Yan X. Application of Pharmacokinetic-Pharmacodynamic Modeling in Drug Delivery: Development and Challenges. Frontiers in Pharmacology 2020;11:997.

[150] Lin J, Lu A. Role of Pharmacokinetics and Metabolism in Drug Discovery and Development. Pharmacological Reviews 1997;49(4):403-449.

[151] Sharma P, Patel N, Prasad B, Varma M. Pharmacokinetics: Theory and Application in Drug Discovery and Development. Drug Discovery and Development 2021:297-355.

[152] Sood R, A. A. Pharmacokinetic and Pharmacodynamic Modeling (PK/PD) in Pharmaceutical Research: Current Research and Advances. Software and Programming Tools in Pharmaceutical Research 2024:153-169.

[153] Ghani S, Khan N, Sable H, Yao F, Shafiq M. Computational techniques for enhancing PK/PD modeling and simulation and ADMET prediction. Computational Methods in Medicinal Chemistry, Pharmacology, and Toxicology 2025:153-174.

[154] Liu X, Liu C, Huang R, Zhu H, Liu Q, Mitra S, Wang Y. Long short-term memory recurrent neural network for pharmacokinetic-pharmacodynamic modeling. Int. Journal of Clinical Pharmacology and Therapeutics 2021;59(02):138-146.

[155] Liu C, Xu Y, Liu Q, Zhu H, Wang Y. Application of machine learning based methods in exposure–response analysis. Journal of Pharmacokinetics and Pharmacodynamics 2022;49(4):401-410.

[156] Tang A. Machine Learning for Pharmacokinetic/Pharmacodynamic Modeling. Journal of Pharmaceutical Sciences 2023;112(5):1460-1475.

[157] Keutzer L, You H, Farnoud A, Nyberg J, Wicha S, Maher-Edwards G, Vlasakakis G, Moghaddam G, Svensson E, Menden M, Simonsson U. Machine Learning and Pharmacometrics for Prediction of Pharmacokinetic Data: Differences, Similarities and Challenges Illustrated with Rifampicin. Pharmaceutics 2022;14(8):1530.

[158] Shardlow C, Generaux G, MacLauchlin C, Pons N, Skordos K, Bloomer J. Utilizing Drug-Drug Interaction Prediction Tools during Drug Development: Enhanced Decision Making Based on Clinical Risk. Drug Metabolism and Disposition 2011;39(11):2076-2084.

[159] Wang N, Zhu B, Li X, Liu S, Shi J, Cao D. Comprehensive Review of Drug–Drug Interaction Prediction Based on Machine Learning: Current Status, Challenges, and Opportunities. Journal of Chemical Information and Modeling 2023;64(1):96-109.

[160] Han K, Cao P, Wang Y, Xie F, Ma J, Yu M, Wang J, Xu Y, Zhang Y, Wan J. A Review of Approaches for Predicting Drug–Drug Interactions Based on Machine Learning. Frontiers in Pharmacology 2022;12:814858.

[161] Ogidi O, Alfred-Ugbenbo D, Poripo B. AI-Driven

Drug Discovery, Design, and Development in Immunological Disorders. Advances in Computational Intelligence and Robotics 2025:143-180.

[162] Patil R. Research and development in the pharmaceutical industry. Research and Development 2024;1(1).

[163] Marques L, Costa B, Pereira M, Silva A, Santos J, Saldanha L, Silva I, Magalhães P, Schmidt S, Vale N. Advancing Precision Medicine: A Review of Innovative In Silico Approaches for Drug Development, Clinical Pharmacology and Personalized Healthcare. Pharmaceutics 2024;16(3):332.

[164] Yin Q, Fan R, Cao X, Liu Q, Jiang R, Zeng W. DeepDrug: A general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction. Quantitative Biology 2023;11(3):260-274.

[165] Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, Friedman C. Drug—drug interaction through molecular structure similarity analysis. Journal of the American Medical Informatics Association 2012;19(6):1066-1074.

[166] Huang J, Niu C, Green C, Yang L, Mei H, Han J. Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network. PLoS Computational Biology 2013;9(3):e1002998.

[167] Cami A, Manzi S, Arnold A, Reis B. Pharmacointeraction Network Models Predict Unknown Drug-Drug Interactions. PLoS ONE 2013;8(4):e61468.

[168] Lawson C, Hodgson J, Wilson R, Richards S. Prevalence, thresholds and the performance of presence–absence models. Methods in Ecology and Evolution 2013;5(1):54-64.

[169] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. Journal of the American Medical Informatics Association 2014;21(e2):e278-e286.

[170] Liu S, Huang Z, Qiu Y, Chen Y, Zhang W. Structural Network Embedding using Multi-modal Deep Auto-encoders for Predicting Drug-drug Interactions. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2019:445-450.

[171] Jang H, Song J, Kim J, Lee H, Kim I, Moon B, Oh J. Machine learning-based quantitative prediction of drug exposure in drug-drug interactions using drug label information. npj Digital Medicine 2022;5(1):88.

[172] Grimes D, Schulz K. An overview of clinical research: the lay of the land. The Lancet 2002;359(9300):57-61.

[173] Nair B. Clinical Trial Designs. Indian Dermatology Online Journal 2019;10(2):193.

[174] Li A, Bergan R. Clinical trial design: Past, present, and future in the context of big data and precision medicine. Cancer 2020;126(22):4838-4846.

[175] Bieske L, Zinner M, Dahlhausen F, Truebel H. Critical path activities in clinical trial setup and conduct: How to avoid bottlenecks and accelerate clinical trials. Drug Discovery Today 2023;28(10):103733.

[176] Council for International Organizations of Medical Sciences et al. Clinical research in resource-limited settings. Council for International Organizations of Medical Sciences 2021.

[177] Feijoo F, Palopoli M, Bernstein J, Siddiqui S, Albright T. Key indicators of phase transition for clinical trials through machine learning. Drug Discovery Today 2020;25(2):414-421.

[178] Wu K, Wu E, DAndrea M, Chitale N, Lim M, Dabrowski M, Kantor K, Rangi H, Liu R, Garmhausen M, Pal N, Harbron C, Rizzo S, Copping R, Zou J. Machine Learning Prediction of Clinical Trial Operational Efficiency. The AAPS Journal 2022;24(3):57.

[179] Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. Trends in Pharmacological Sciences 2019;40(8):577-591.

[180] Kimko HC, Duffull SB. Simulation for designing clinical trials. Marcel Dekker Incorporated 2002.

[181] Follett L, Geletta S, Laugerman M. Quantifying risk associated with clinical trial termination: A text mining approach. Information Processing & Management 2019;56(3):516-525.

[182] Kavalci E, Hartshorn A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. Scientific Reports 2023;13(1):121.

[183] Elkin M, Zhu X. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. Scientific Reports 2021;11(1):3446.

[184] Gresham G. ClinicalTrials.gov. Principles and Practice of Clinical Trials 2020:1-18.

[185] Chalmers E, Hill D, Zhao V, Lou E. Prescriptive analytics applied to brace treatment for AIS: a pilot demonstration. Scoliosis 2015;10(S2):1-4.

[186] Ezike T, Okpala U, Onoja U, Nwike C, Ezeako E, Okpara O, Okoroafor C, Eze S, Kalu O, Odoh E, Nwadike U, Ogbodo J, Umeh B, Ossai E, Nwanguma B. Advances in drug delivery systems, challenges and future directions. Heliyon 2023;9(6):e17488.

[187] Homayun B, Lin X, Choi H. Challenges and Recent Progress in Oral Drug Delivery Systems for Biopharmaceuticals. Pharmaceutics 2019;11(3):129.

[188] Ugurlu S. Investigation of metallacages for cisplatin encapsulation using Density Functional Theory (DFT). OAJ Materials and Devices 2024;8

[189] Mitchell M, Billingsley M, Haley R, Wechsler M, Peppas N, Langer R. Engineering precision nanoparticles for drug delivery. Nature Reviews Drug Discovery 2020;20(2):101-124.

[190] Wen H, Jung H, Li X. Drug Delivery Approaches in Addressing Clinical Pharmacology-Related Issues: Opportunities and Challenges. The AAPS Journal 2015;17(6):1327-1340.

[191] Bhandare A, Nannor KM. Bioavailability in drug design and development: A comprehensive review. World Journal of Pharmaceutical Research 2024;13(17):145–168.

[192] Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal 2021;8(2):e188-e194.

[193] Vidhya K, Sultana A, M N, Rangareddy H. Artificial Intelligence's Impact on Drug Discovery and Development From Bench to Bedside. Cureus 2023;15(10).

[194] Vora L, Gholap A, Jetha K, Thakur R, Solanki H, Chavda V. Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. Pharmaceutics 2023;15(7):1916.

[195] Deng J, Dong W, Socher R, Li L, Kai Li , Li Fei-Fei . ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009:248-255.

[196] Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Advanced Drug Delivery Reviews 2021;175:113806.

[197] Mhatre S, Shukla S, Chavda V, Gandikota L, Patravale V. AI and ML for Development of Cell and Gene Therapy for Personalized Treatment. Bioinformatics Tools for Pharmaceutical Drug Product Development 2023:371-400.

[198] Dong Y, Yang T, Xing Y, Du J, Meng Q. Data-Driven Modeling Methods and Techniques for Pharmaceutical Processes. Processes 2023;11(7):2096.

[199] Bhattamisra S, Banerjee P, Gupta P, Mayuren J, Patra S, Candasamy M. Artificial Intelligence in Pharmaceutical and Healthcare Research. Big Data and Cognitive Computing 2023;7(1):10.

[200] Minich D, Bland J. Personalized Lifestyle Medicine: Relevance for Nutrition and Lifestyle Recommendations. The Scientific World Journal 2013;2013(1):129841.

[201] Larry Jameson J, Longo D. Precision Medicine—Personalized, Problematic, and Promising. Obstetrical & Gynecological Survey 2015;70(10):612-614.

[202] Gray I, Kross A, Renfrew M, Wood P. Precision Medicine in Lifestyle Medicine: The Way of the Future?. American Journal of Lifestyle Medicine 2019;14(2):169-186.

[203] KRAVITZ R, DUAN N, BRASLOW J. Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages. The Milbank Quarterly 2004;82(4):661-687.

[204] Hartman E, Grieve R, Ramsahai R, Sekhon J. From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects. Journal of the Royal Statistical Society Series A: Statistics in Society 2015;178(3):757-778.

[205] Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database 2020:baaa010.

[206] Rahman MH, Islam T, Hossen ME, Chowdhury ME, Hayat R. Machine learning in healthcare: From diagnostics to personalized medicine and predictive analytics. Integrative Biomedical Research 2024;8(12):1–8.

[207] Libbrecht M, Noble W. Machine learning applications in genetics and genomics. Nature Reviews Genetics 2015;16(6):321-332.

[208] Zhang Y, Li G, Bian W, Bai Y, He S, Liu Y, Liu H, Liu J. Value of genomics- and radiomics-based machine learning models in the identification of breast cancer molecular subtypes: a systematic review and meta-analysis. Annals of Translational Medicine 2022;10(24):1394-1394.

[209] Abraham J, Heimberger A, Marshall J, Heath E, Drabick J, Helmstetter A, Xiu J, Magee D, Stafford P, Nabhan C, Antani S, Johnston C, Oberley M, Korn W, Spetzler D. Machine learning analysis using 77,044 genomic and transcriptomic profiles to accurately predict tumor type. Translational Oncology 2021;14(3):101016.

[210] Afrifa-Yamoah E, Adua E, Peprah-Yamoah E, Anto E, Opoku-Yamoah V, Acheampong E, Macartney M, Hashmi R. Pathways to chronic disease detection and prediction: Mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges. Chronic Diseases and Translational Medicine 2024;11(1):1-21.

[211] Kaur Thethi S. Chapter 8 Machine learning models for cost-effective healthcare delivery systems: A global perspective. Digital Transformation in Healthcare 5.0 2024;:199-244.

[212] Topol E. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 2019;25(1):44-56.

[213] Chakilam C. Next-generation healthcare: Merging ai, ml, and big data for accelerated disease diagnosis and personalized treatment. American Online Journal of Science and Engineering (AOJSE) 2023;1(1).

[214] Marcu LG, Marcu DC. Pharmacogenomics and big data in medical oncology: developments and challenges. Therapeutic advances in medical oncology 2024;16:17588359241287658.

[215] Wang H, Huang J, Fang X, Liu M, Fan X, Li Y. Advances in next-generation sequencing (NGS) applications in drug discovery and development. Expert Opinion on Drug Discovery 2025;20(4):537-550.

[216] Barros M, Paci M, Tervonen A, Passini E, Koivumäki J, Hyttinen J, Lenk K. From Multiscale Biophysics to Digital Twins of Tissues and Organs: Future Opportunities for in-silico Pharmacology. IEEE Transactions on Molecular, Biological, and Multi-Scale Communications 2024;10(4):576-594.

[217] Corti A, Colombo M, Migliavacca F, Rodriguez Matas J, Casarin S, Chiastra C. Multiscale Computational Modeling of Vascular Adaptation: A Systems Biology Approach Using Agent-Based Models. Frontiers in Bioengineering and Biotechnology 2021;9:744560.

[218] Gradeci D, Bove A, Charras G, Lowe A, Banerjee S. Single-cell approaches to cell competition: High-throughput imaging, machine learning and simulations. Seminars in Cancer Biology 2020;63:60-68.

[219] Sivakumar N, Mura C, Peirce S. Innovations in integrating machine learning and agent-based modeling of biomedical systems. Frontiers in Systems Biology 2022;2:959665.

[220] Cogno N, Axenie C, Bauer R, Vavourakis V. Agent-based modeling in cancer biomedicine: applications and tools for calibration and validation. Cancer Biology & Therapy 2024;25(1):2344600.

[221] Deist T, Patti A, Wang Z, Krane D, Sorenson T, Craft D. Simulation-assisted machine learning. Bioinformatics 2019;35(20):4072-4080.

[222] Ali M, Benfante V, Basirinia G, Alongi P, Sperandeo A, Quattrocchi A, Giannone A, Cabibi D, Yezzi A, Di Raimondo D, Tuttolomondo A, Comelli A. Applications of Artificial Intelligence, Deep Learning, and Machine Learning to Support the Analysis of Microscopic Images of Cells and Tissues. Journal of Imaging 2025;11(2):59.

[223] Choi H, Wang C, Pan X, Jang J, Cao M, Brazzo J, Bae Y, Lee K. Emerging machine learning approaches to phenotyping cellular motility and morphodynamics. Physical Biology 2021;18(4):041001.

[224] Lee R, Wu Y, Goh D, Tan V, Ng C, Lim J, Lau M, Yeong J. Application of Artificial Intelligence to In Vitro Tumor Modeling and Characterization of the Tumor Microenvironment. Advanced Healthcare Materials 2023;12(14):2202457.

[225] Wang S, Rong R, Yang DM, Fujimoto J, Yan S, Cai L, Yang L, Luo D, Behrens C, Parra ER. Computational staining of pathology images to study the tumor microenvironment in lung cancer. Cancer research 2020;80(10):2056–2066.

[226] He W, Kong S, Lin R, Xie Y, Zheng S, Yin Z, Huang X, Su L, Zhang X. Machine Learning Assists in the Design and Application of Microneedles. Biomimetics 2024;9(8):469.

[227] Datta S, Islam M, Rahman Sobuz M, Ahmed S, Kar M. Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: A comprehensive review. Heliyon 2024;10(5):e26888.

[228] Clancy C, An G, Cannon W, Liu Y, May E, Ortoleva P, Popel A, Sluka J, Su J, Vicini P, Zhou X, Eckmann D. Multiscale Modeling in the Clinic: Drug Design and Development. Annals of Biomedical Engineering 2016;44(9):2591-2610.

[229] Gondal M, Chaudhary S. Navigating Multi-Scale Cancer Systems Biology Towards Model-Driven Clinical Oncology and Its Applications in Personalized Therapeutics. Frontiers in Oncology 2021;11:712505.

[230] Kim S, Thiessen P, Bolton E, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker B, Wang J, Yu B, Zhang J, Bryant S. PubChem Substance and Compound databases. Nucleic Acids Research 2015;44(D1):D1202-D1213.

[231] Gaulton A, Hersey A, Nowotka M, Bento A, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis L, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños M, Overington J, Papadatos G, Smit I, Leach A. The ChEMBL database in 2017. Nucleic Acids Research 2016;45(D1):D945-D954.

[232] Chen X, Liu M, Gilson M. BindingDB: A Web-Accessible Molecular Recognition Database. Combinatorial Chemistry & High Throughput Screening 2001;4(8):719-725.

[233] Berman H, Battistuz T, Bhat T, Bluhm W, Bourne P, Burkhardt K, Feng Z, Gilliland G, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook J, Zardecki C. The Protein Data Bank. Acta Crystallographica Section D Biological Crystallography 2002;58(6):899-907.

[234] Wang R, Fang X, Lu Y, Yang C, Wang S. The PDBbind Database: Methodologies and Updates. Journal of Medicinal Chemistry 2005;48(12):4111-4119.

[235] Irwin J, Shoichet B. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. Journal of Chemical Information and Modeling 2004;45(1):177-182.

[236] Mysinger M, Carchia M, Irwin J, Shoichet B. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. Journal of Medicinal Chemistry 2012;55(14):6582-6594.

[237] Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, Wang R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. Journal of Chemical Information and Modeling 2018;59(2):895-913.

[238] Ruddigkeit L, Blum L, Reymond J. Visualization and Virtual Screening of the Chemical Universe Database

GDB-17. Journal of Chemical Information and Modeling 2013;53(1):56-65.

[239] Knox C, Wilson M, Klinger C, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin N, Strawbridge S, Garcia-Patino M, Kruger R, Sivakumaran A, Sanford S, Doshi R, Khetarpal N, Fatokun O, Doucet D, Zubkowski A, Rayat D, Jackson H, Harford K, Anjum A, Zakir M, Wang F, Tian S, Lee B, Liigand J, Peters H, Wang R, Nguyen T, So D, Sharp M, da Silva R, Gabriel C, Scantlebury J, Jasinski M, Ackerman D, Jewison T, Sajed T, Gautam V, Wishart D. DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Research 2023;52(D1):D1265-D1275.

[240] Rudolph J, Settleman J, Malek S. Emerging Trends in Cancer Drug Discovery—From Drugging the "Undruggable" to Overcoming Resistance. Cancer Discovery 2021;11(4):815-821.

[241] Cox A, Fesik S, Kimmelman A, Luo J, Der C. Drugging the undruggable RAS: Mission Possible?. Nature Reviews Drug Discovery 2014;13(11):828-851.

[242] Ostrem J, Peters U, Sos M, Wells J, Shokat K. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature 2013;503(7477):548-551.

[243] Berndt N, Hamilton A, Sebti S. Targeting protein prenylation for cancer therapy. Nature Reviews Cancer 2011;11(11):775-791.

[244] Lazo J, Sharlow E. Drugging Undruggable Molecular Cancer Targets. Annual Review of Pharmacology and Toxicology 2016;56(1):23-40.

[245] Backus K, Correia B, Lum K, Forli S, Horning B, González-Páez G, Chatterjee S, Lanning B, Teijaro J, Olson A, Wolan D, Cravatt B. Proteome-wide covalent ligand discovery in native biological systems. Nature 2016;534(7608):570-574.

[246] Boike L, Cioffi A, Majewski F, Co J, Henning N, Jones M, Liu G, McKenna J, Tallarico J, Schirle M, Nomura D. Discovery of a Functional Covalent Ligand Targeting an Intrinsically Disordered Cysteine within MYC. Cell Chemical Biology 2021;28(1):4-13.

[247] Sievers Q, Petzold G, Bunker R, Renneville A, Słabicki M, Liddicoat B, Abdulrahman W, Mikkelsen T, Ebert B, Thomä N. Defining the human C2H2 zinc finger degrome targeted by thalidomide analogs through CRBN. Science 2018;362(6414):eaat0572.

[248] Du X, Volkov O, Czerwinski R, Tan H, Huerta C, Morton E, Rizzi J, Wehn P, Xu R, Nijhawan D, Wallace E. Structural Basis and Kinetic Pathway of RBM39 Recruitment to DCAF15 by a Sulfonamide Molecular Glue E7820. Structure 2019;27(11):1625-1633.

[249] Samarasinghe K, Crews C. Targeted protein degradation: A promise for undruggable proteins. Cell Chemical Biology 2021;28(7):934-951.

[250] Ugurlu S, McDonald D, Enisoglu R, Zhu Z, He S. MEGA PROTAC, MEGA DOCK-based PROTAC mediated ternary complex formation pipeline with sequential filtering and rank aggregation. Scientific Reports 2025;15(1):5545.

[251] Petersen D, Hawkins J, Ruangsiriluk W, Stevens K, Maguire B, O'Connell T, Rocke B, Boehm M, Ruggeri R, Rolph T, Hepworth D, Loria P, Carpino P. A Small-Molecule Anti-secretagogue of PCSK9 Targets the 80S Ribosome to Inhibit PCSK9 Protein Translation. Cell Chemical Biology 2016;23(11):1362-1371.

[252] Warner K, Hajdin C, Weeks K. Principles for targeting RNA with drug-like small molecules. Nature Reviews Drug Discovery 2018;17(8):547-558.

[253] Zhuo C, Gao J, Li A, Liu X, Zhao Y. A Machine Learning Method for RNA–Small Molecule Binding Preference Prediction. Journal of Chemical Information and Modeling 2024;64(19):7386-7397.

[254] Xia W, Shu J, Sang C, Wang K, Wang Y, Sun T, Xu X. The prediction of RNA-small-molecule ligand binding affinity based on geometric deep learning. Computational Biology and Chemistry 2025;115:108367.

[255] Wang K, Zhou R, Wu Y, Li M. RLBind: a deep learning method to predict RNA–ligand binding sites. Briefings in Bioinformatics 2022;24(1):bbac486.

[256] Li J, Tan Y, Lu R, Liang P, Liu H, Yao X. Artificial intelligence for RNA–ligand interaction prediction: advances and prospects. Drug Discovery Today 2025;30(6):104366.

[257] Gui J, Chen T, Zhang J, Cao Q, Sun Z, Luo H, Tao D. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. IEEE Transactions on Pattern Analysis and Machine Intelligence 2024;46(12):9052-9071.

[258] Rani V, Nabi S, Kumar M, Mittal A, Kumar K. Self-supervised Learning: A Succinct Review. Archives of Computational Methods in Engineering 2023;30(4):2761-2775.

[259] Chen Y, Wang Z, Wang J, Chu Y, Zhang Q, Li Z, Zeng X. Self-supervised learning in drug discovery. Science China Information Sciences 2025;68(7):170103.

[260] Rani V, Kumar M, Gupta A, Sachdeva M, Mittal A, Kumar K. Self-supervised learning for medical image analysis: a comprehensive review. Evolving Systems 2024;15(4):1607-1633.

[261] Li Z, Jiang M, Wang S, Zhang S. Deep learning methods for moleculr representation and property prediction. Drug Discovery Today 2022;27(12):103373.

[262] Mswahili M, Jeong Y. Transformer-based models for chemical SMILES representation: A comprehensive literature review. Heliyon 2024;10(20):e39038.

[263] Sultan A, Sieg J, Mathea M, Volkamer A. Transformers for Molecular Property Prediction: Lessons Learned from the Past Five Years. Journal of Chemical Information and Modeling 2024;64(16):6259-6280.

[264] Das M, Ghosh A, Sunoj R. Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. Journal of Computational Chemistry 2024;45(14):1160-1176.

[265] Chithrananda S, Grand G, Ramsundar B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint 2020:2010.09885.

[266] Miao Y, Ma H, Huang J. Recent Advances in Toxicity Prediction: Applications of Deep Graph Learning. Chemical Research in Toxicology 2023;36(8):1206-1226.

[267] Wang H, Kaddour J, Liu S, Tang J, Lasenby J, Liu Q. Evaluating self-supervised learning for molecular graph embeddings. Advances in Neural Information Processing Systems 2023;36:68028–68060.

[268] Jiang X, Tan L, Zou Q. DGCL: dual-graph neural networks contrastive learning for molecular property prediction. Briefings in Bioinformatics 2024;25(6).

[269] Qian Y, Zhang Y, Ye Y, Zhang C. Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media. Advances in Neural Information Processing Systems 2021;34:26911–26923.

[270] Meng Z, Li Y, Zhao P, Yu Y, King I. Meta-Learning with Motif-based Task Augmentation for Few-Shot Molecular Property Prediction. Proceedings of the 2023 SIAM International Conference on Data Mining (SDM) 2023:811-819.

[271] Vettoruzzo A, Bouguelia M, Vanschoren J, Rögnvaldsson T, Santosh K. Advances and Challenges in Meta-Learning: A Technical Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 2024;46(7):4763-4779.

[272] Francisca Chibugo Udegbe , Ogochukwu Roseline Ebulue , Charles Chukwudalu Ebulue , Chukwunonso Sylvester Ekesiobi . MACHINE LEARNING IN DRUG DISCOVERY: A CRITICAL REVIEW OF APPLICATIONS AND CHALLENGES. Computer Science & IT Research Journal 2024;5(4):892-902.

[273] Dara S, Dhamercherla S, Jadav S, Babu C, Ahsan M. Machine Learning in Drug Discovery: A Review. Artificial Intelligence Review 2021;55(3):1947-1999.

[274] Chandak T, Jayashree J, Vijayashree J. Trends and Advancements of AI and XAI in Drug Discovery. Explainable AI (XAI) for Sustainable Development 2024:233-255.

[275] Kanti P, Prabhakar Sharma , Wanatasanappan V, Said N. Explainable machine learning techniques for hybrid nanofluids transport characteristics: an evaluation of shapley additive and local interpretable model-agnostic explanations. Journal of Thermal Analysis and Calorimetry 2024;149(21):11599-11618.

[276] Hooshyar D, Yang Y. Problems With SHAP and LIME in Interpretable AI for Education: A Comparative

Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction. IEEE Access 2024;12:137472-137490.

[277] Li S, Wang X, Barnard A. Diverse explanations from data-driven and domain-driven perspectives in the physical sciences. Machine Learning: Science and Technology 2025;6(1):013002.

[278] Shahin R, Jaafreh S, Azzam Y. Tracking protein kinase targeting advances: integrating QSAR into machine learning for kinase-targeted drug discovery. Future Science OA 2025;11(1):2483631.

[279] Ma X, Wang J, Ni X, Shi J. Machine Learning Approaches for Enhancing Customer Retention and Sales Forecasting in the Biopharmaceutical Industry: A Case Study. Journal of Advanced Computing Systems 2024;4(1):10-29.

[280] Subramanian MK, Spencer J. Deep neural networks for in silico prediction of drug-target interactions: Advancements and future directions. Frontiers in Emerging Artificial Intelligence and Machine Learning 2025;2(03):1–12.

[281] Chowdhury S, Sany M, Ahamed M, Das S, Badal F, Das P, Tasneem Z, Hasan M, Islam M, Ali M, Abhi S, Islam M, Sarker S. A State-of-the-Art Computer Vision Adopting Non-Euclidean Deep-Learning Models. International Journal of Intelligent Systems 2023;2023(1):8674641.

[282] Liu M, Li C, Chen R, Cao D, Zeng X. Geometric deep learning for drug discovery. Expert Systems with Applications 2024;240:122498.

[283] Isert C, Atz K, Schneider G. Structure-based drug design with geometric deep learning. Current Opinion in Structural Biology 2023;79:102548.

[284] Liao Z, You R, Huang X, Yao X, Huang T, Zhu S. DeepDock: Enhancing Ligand-protein Interaction Prediction by a Combination of Ligand and Structure Information. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2019:311-317.

[285] Stark H, Ganea O, Pattanaik L, Barzilay R, Jaakkola T. Equibind: Geometric deep learning for drug binding structure prediction. In International conference on machine learning. PMLR 2022:20503–20521.

[286] Corso G, Stark H, Jing B, Barzilay R, Jaakkola T. Diffdock: Diffusion steps, twists, and turns for molecular docking. arXiv preprint 2022:2210.01776.

[287] Zhang X, Shen C, Zhang H, Kang Y, Hsieh C, Hou T. Advancing Ligand Docking through Deep Learning: Challenges and Prospects in Virtual Screening. Accounts of Chemical Research 2024;57(10):1500-1509.

[288] Garc´ıa-Vinuesa J, Rojas J, Soto-Garc´ıa N, Mart´ınez N. Geometric deep learning assists protein engineering opportunities and challenges. arXiv preprint 2025:2506.16091.

[289] Kumar P. Diffusion Models and Generative Artificial Intelligence: Frameworks, Applications and Challenges. Archives of Computational Methods in Engineering 2025:1-44.

[290] Schneuing A, Harris C, Du Y, Didi K, Jamasb A, Igashov I, Du W, Gomes C, Blundell T, Lio P, Welling M, Bronstein M, Correia B. Structure-based drug design with equivariant diffusion models. Nature Computational Science 2024;4(12):899-909.

[291] Alakhdar A, Poczos B, Washburn N. Diffusion Models in De Novo Drug Design. Journal of Chemical Information and Modeling 2024;64(19):7238-7256.

[292] Xu M, Yu L, Song Y, Shi C, Ermon S, Tang J. Geodiff: A geometric diffusion model for molecular conformation generation. arXiv preprint 2022:2203.02923.

[293] Guo Z, Liu J, Wang Y, Chen M, Wang D, Xu D, Cheng J. Diffusion models in bioinformatics and computational biology. Nature Reviews Bioengineering 2023;2(2):136-154.

[294] Chen Z, Peng B, Zhai T, Adu-Ampratwum D, Ning X. Generating 3d binding molecules using shape-conditioned diffusion models with guidance. arXiv preprint 2025:2502.06027.

[295] Chen Z, Xu L, Zheng H, Chen L, Tolba A, Zhao L, Yu K, Feng H. Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. Computers, Materials & Continua 2024;80(2):1753-1808.

[296] Huang J, Xu Y, Wang Q, Wang Q, Liang X, Wang F, Zhang Z, Wei W, Zhang B, Huang L, Chang J, Ma L, Ma T, Liang Y, Zhang J, Guo J, Jiang X, Fan X, An Z, Li T, Li X, Shao Z, Qian T, Sun T, Diao B, Yang C, Yu C, Wu Y, Li M, Zhang H, Zeng Y, Zhang Z, Zhu Z, Lv Y, Li A, Chen X, An B, Xiao W, Bai C, Mao Y, Yin Z, Gui S, Su W, Zhu Y, Gao J, He X, Li Y, Jin G, Ao X, Zhai X, Tan H, Yun L, Shi H, Li J, Fan C, Huang K, Harrison E, Leung V, Qiu S, Dong Y, Zheng X, Wang G, Zheng Y, Wang Y, Guo J, Wang L, Cheng X, Wang Y, Yang S, Fu M, Fei A. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. The Innovation 2025;6(6):100948.

[297] Zheng Y, Koh HY, Yang M, Li L, May LT, Webb GI, Pan S, Church G. Large language models in drug discovery and development: From disease mechanisms to clinical trials. arXiv preprint 2024:2409.04481.

[298] Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022;38(8):2102-2110.

[299] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 2022;44(10):7112-7127.

[300] Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, Poulton A, Kerkez V, Stojnic R. Galactica: A large language model for science. arXiv preprint 2022:2211.09085.

[301] Lee H, Kim J, Kim J, Lee Y. Recent advances in ai-based toxicity prediction for drug discovery. Frontiers in Chemistry, 2025;13:1632046.

[302] Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nature Reviews Drug Discovery 2023;23(2):141-155.

[303] Ren Q, Qu N, Sun J, Liu J, Ni L, Tong X, Zhang Z, Kong X, Wen Y, Wang Y, Wang D, Luo X, Zheng M, Li X. KinomeMETA: meta-learning enhanced kinome-wide polypharmacology profiling. Briefings in Bioinformatics 2023;25(1):bbad461.

[304] Mukherjee A, Abraham S, Singh A, Balaji S, Mukunthan K. From Data to Cure: A Comprehensive Exploration of Multi-omics Data Analysis for Targeted Therapies. Molecular Biotechnology 2024;67(4):1269-1289.

[305] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinformatics and Biology Insights 2020;14:117793221989905.

[306] Nguyen N, Wang D. Multiview learning for understanding functional multiomics. PLOS Computational Biology 2020;16(4):e1007677.

[307] Vahabi N, Michailidis G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. Frontiers in Genetics 2022;13:854752.

[308] Turanli B, Karagoz K, Gulfidan G, Sinha R, Mardinoglu A, Arga K. A Network-Based Cancer Drug Discovery: From Integrated Multi-Omics Approaches to Precision Medicine. Current Pharmaceutical Design 2019;24(32):3778-3790.

[309] Hassan Ali . Artificial intelligence in multi-omics data integration: Advancing precision medicine, biomarker discovery and genomic-driven disease interventions. International Journal of Science and Research Archive 2023;8(1):1012-1030.

[310] She S, Chen H, Ji W, Sun M, Cheng J, Rui M, Feng C. Deep learning-based multi-drug synergy prediction model for individually tailored anti-cancer therapies. Frontiers in Pharmacology 2022;13:1032875.