Machine Learning Applications in Drug Discovery

Sadettin Y. Ugurlu

Department of Computer Science, University of Birmingham, the UK

s.yavuz.ugurlu@gmail.com

Abstract

Integrating machine learning (ML) into drug discovery has ushered in a new era of innovation, dramatically enhancing the efficiency and precision of identifying and developing new therapeutics. This review provides a comprehensive analysis of the current applications of machine learning in drug discovery, focusing on its transformative impact across various stages of the drug development pipeline. We delve into key ML methodologies, including supervised and unsupervised learning, neural networks, and reinforcement learning, examining their underlying principles and specific contributions to drug discovery processes. By exploring case studies and recent advancements, this review illustrates how ML algorithms have been utilized to predict drug-target interactions, optimize drug design, and streamline clinical trial processes. Furthermore, we discuss the challenges and limitations of implementing ML techniques in this field and highlight emerging trends and future directions. This review aims to offer researchers a thorough understanding of ML's potential to revolutionize drug discovery and equip them with the insights needed to leverage these technologies effectively.

Keywords: Drug discovery; drug development; machine learning; artifical integellence; deep learning

1 Introduction

Artificial Intelligence, sometimes called machine intelligence, is a concept that pertains to the ability of a computer system to acquire knowledge and skills through the analysis of past data ([1]). Machine Learning (ML) techniques have been widely employed across various domains, including but not limited to assisted autonomous vehicles, improved speech recognition, medical applications, natural language processing and drug discovery and development ([2]).

Drug discovery and development pipelines are extensive, intricate, and contingent upon many factors ([3]). ML methodologies offer a range of tools that have the potential to enhance the process of exploration and decision-making for precisely defined inquiries, utilising ample and reliable data. Machine learning (ML) can be applied at several stages of the drug discovery process. Before implementing ML approaches, it is crucial to have a profound understanding of the foundational principles of Machine Learning. Therefore, the application of ML in the various stages of drug development will be examined in the subsequent sixteen subsections.

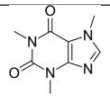
1.1 Virtual Screening Predictions

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 ([4]) (Table 1. Whereas receptor-based methods use molecular docking to forecast binding affinities and interactions, ligand-based techniques examine molecular similarity and structure-activity connections. These techniques, taken together, improve the accuracy and scope of virtual screening, therefore hastening the identification of possible drug candidates with specific pharmacological effects ([5]).

One must have a representation of the drug candidate and targets if one is training a machine learning model utilising ligand and receptor (target) information ([6]). Ligand-based drug design is the machine learning-based analysis of molecular characteristics and chemical fingerprints of active and inactive compounds ([7]). By matching newly discovered compounds to known active molecules, ML systems can use patterns found from this data to forecast their activity. This approach is practical when there is a shortage of particular and thorough structural data about the receptor being targeted or when one wants to target several receptors with similar binding sites for ligands ([7]).

Conversely, receptor-based drug design uses molecular docking simulations with machine learning. Under these conditions, machine learning techniques use simulations to predict the degree of attraction between a small molecule ligand and a target receptor from their spatial and electrostatic compatibility. These models help to identify potential therapy candidates with excellent specificity and potency by using known ligand-receptor interactions and structural data to forecast the most favourable binding locations and strengths ([8]).

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 sequenced based feature.	CC(=O) NC1 = CC = C(C = C1) O
	Fingerprint	It represent the molecule structure that convert into bit string i.e., presence of atoms indicate 1 and absence of atom indicate 0; it is useful method for describing the structural similarity of the molecule.	(1,1,0,0. . . . . 0,1,0,1,1)
Structure-Based Feature	Molecular Graph	Molecular graph represents the chemical structure in terms of graph theory. It is the mapping of atoms constituting a molecule to node and chemical bonds to edges.	
	Voxel	Voxel is the combination of volume and pixel representation of three dimensions space. It is used in the target protein because it reacts with ligand instead of the entire protein and it is very suitable for binding prediction.	3D target protein in cubes.

The different methods for displaying molecular data routinely used in computational approaches for drug discovery and molecular modelling are briefly summarised in this table. Among the representations are atomic coordinates, chemical descriptions, fingerprints, graph-based models, and other specialised forms. Every representation has different advantages and is selected depending on the particular needs of the computing activity. The table provides a thorough overview of the characteristics and applications of every data representation method, helping researchers select the most suitable approach for their computational study ([9]).

By utilising 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 utilising an extensive database comprising chemical compounds and their corresponding efficacy against various bacterial strains ([10]). The development of a generative tensorial reinforcement learning (GENTRL) model by Insilico Medicine resulted in the rapid identification of new highly effective DDR1 kinase inhibitors ([11]). Machine learning techniques have been utilised in several applications, such as exploring potential drugs for Alzheimer's disease ([12]). The research on Alzheimer's disease showcased the capacity of machine learning-based virtual screening approaches for drug discovery.

1.2 Toxicity predictions

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 ([13]). 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 ([14]). For example, clinical trials have revealed that uncontrollable toxicity accounts for 90% of medication development failures ([15]). 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 ([16]). In the given context, the utilisation of computational toxicity prediction proves to be highly beneficial during the initial phases of drug exploration, as it enables

the exclusion of compounds that possess a substantial likelihood of experiencing failure in clinical trials ([14]).

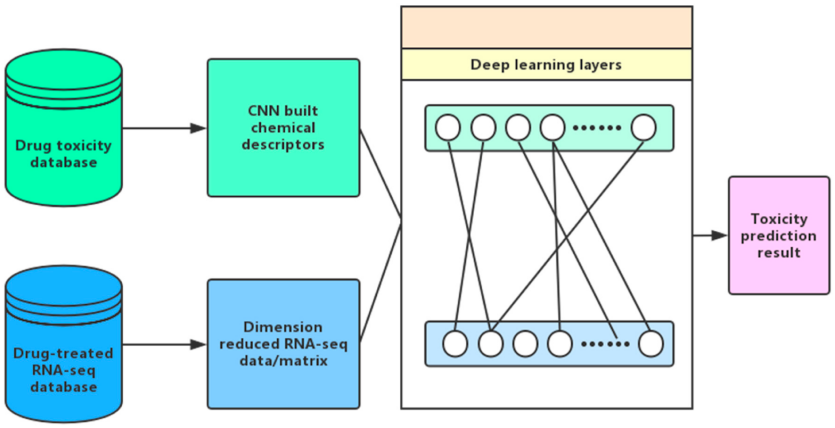


Figure 1: An example of ML application in RNA-seq data's toxicity prediction

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 ([17]).

Quantitative structure-activity relationship (QSAR) models have long been used to predict pharmacological promiscuity and toxicity ([18]). These models show that the biological characteristics of these substances are related to different functional groupings. The spread of toxicity databases in the current setting has helped to adopt machine learning (ML) algorithms to predict the toxicity of small molecules. As such, these models have become somewhat popular recently. Consequently, the scholarly literature has recorded many toxicity prediction methods based on machine learning, including MolToxPred ([19]) and DeepTox ([20]), a neural network model. Although the examples demonstrate the ML models provide promising performance in toxicity predictions, they still have limitations, such as low interpretability.

Despite the numerous advantages of machine learning models in identifying toxicity for a drug candidate, certain limits still need to be addressed. The limitations are:

- Interpretability in machine learning models, particularly deep learning-based ones, poses a challenge ([21, 14].
- The reliability of capturing clinical toxicity in ML models trained on in vitro and in vivo data may be questionable ([21]).
- Although toxicity is a complex issue, most machine-learning models have focused on predicting toxicity in individual platforms using single-task models ([21]).

1.3 Target predictions

Modern drug development depends critically on target predictions since they identify particular biological molecules or cellular processes that drugs can reasonably control to produce therapeutic benefits. Despite the developments in omic and experimental technologies over recent years, the challenge of pinpointing suitable therapeutic targets still

exists ([22]). Target predictions have lately attracted attention as a possible approach by combining multi-omic data with artificial intelligence systems ([23]).

Novel biological targets for small compounds have been predicted via pharmacophore screening ([24]), structure similarity assessment ([25, 26]), reverse docking ([27]) and ML ([22]), depending on the availability of protein structure and the chemical structure of the drug of interest. Despite the rest of the methods, ML is an essential element of artificial intelligence that can be utilised with or without oversight. Supervised learning uses annotated datasets to train models to classify data and accurately predict outcomes. Unsupervised learning, in contrast, investigates the concealed organisation of unlabeled data without human involvement ([28]) (Figure 2).

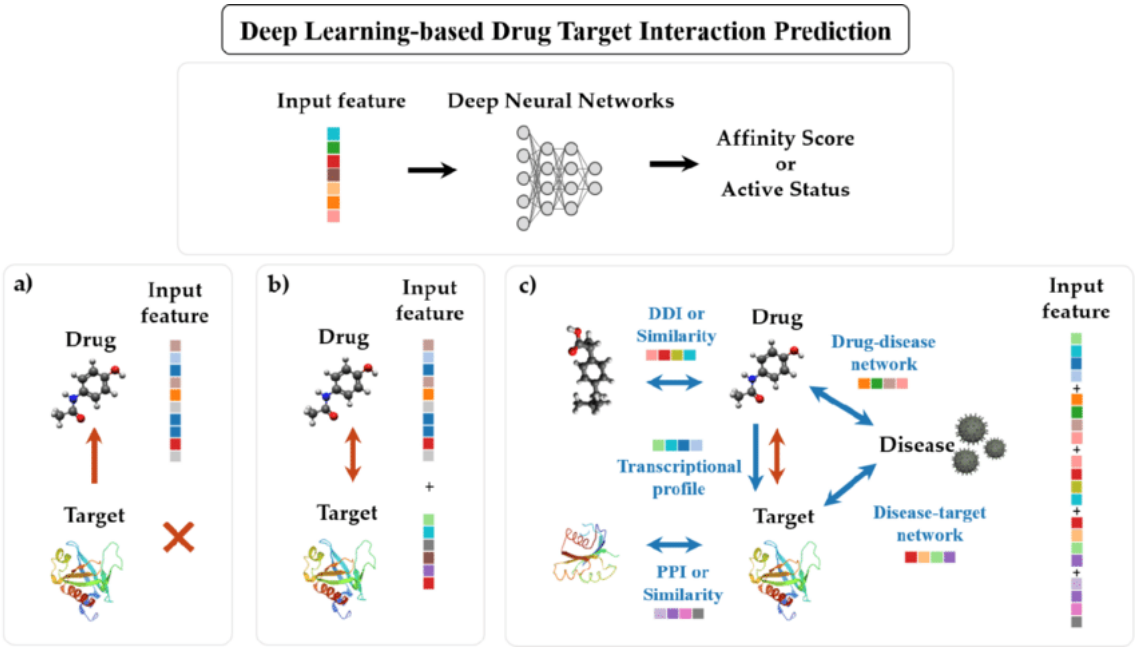


Figure 2: The representation of combining machine learning with pharmaceutical research, deep learning-based drug-target interaction (DTI) prediction

It is a newly developing field. These new methods significantly change the drug development process by using advanced deep learning algorithms to forecast the interactions between drugs and their biological targets. Based on their input features, three branches can be distinguished between DL-based DTI prediction methods: (a) ligand-based approaches, which depend on molecular descriptors and chemical structures of drug molecules; (b) structure-based approaches, which use the 3D structures of target proteins and ligands; and (c) relationship-based approaches, which consider the links between drugs and targets in a network or graph-based representation. Every field offers different advantages and knowledge of drug-target interactions, which helps to develop less dangerous and more effective treatments ([29]).

With the utilisation of ML in the analysis and resolution of intricate biological data networks, such as (Figure 2), ML algorithms have the potential to unveil latent patterns and correlations within the data that may elude human perception. Therefore, ML can potentially enhance comprehension and therapeutic approaches for various diseases. ML has significantly advanced target prediction. Besides target prediction, ML has demonstrated its effectiveness in indication prioritisation, biomarker and target discovery, pharmacokinetics prediction, drug-like molecule creation, clinical trial design, and drug-target interaction. The examples demonstrate that ML has great potential not only in target pre-

dictions but also for other steps in drug discovery.

1.4 Binding affinity predictions

Structure-based drug design (SBDD) is widely used in the search for possible medicinal agents. Usually referred to as ligands, the procedure involves evaluating the docking posture and calculating the degree of interaction between target proteins and small molecules ([30]). Generally referring to binding affinity, the computation of interaction strength uses several scoring systems. Target protein physiological activity is directly proportional to the intensity of their interactions in relation to ligands. Therapeutic candidates thus are selected depending on their capacity to firmly attach to the target protein ([31]). Accurate binding affinity prediction with the help of ML models can decrease the cost of a de novo drug design by utilising the projected binding affinity of the ligand in a library for virtual screening or lead optimisation ([32]). To achieve this objective, conventional scoring functions have been suggested, such as empirical scoring functions ([33]), knowledge-based ([34]), and force-field-based ([35]); typically, these characteristics encompass hydrogen bonds involving desolvation, van der Waals forces (view), and hydrophobic interactions. The coefficient of each descriptor is determined by considering the established binding affinity of protein-ligand complexes ([36]). ML-based scoring functions have been developed to enhance the performance of conventional scoring functions.

Machine learning-based scoring systems have been proposed as substitutes to improve the prediction performance of conventional scoring systems. RF-Score is a prominent machine learning-based grading system ([37]). 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 ([38]) 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 ([39]), it shows better prediction accuracy than the original model. As described in reference ([40]), the Structural Interaction Fingerprints (SIFt) method is a machine learning technique using a format akin to fingerprints to document intermolecular interactions.

1.5 Chemical reaction optimization with ML

Drug discovery and development have significantly benefited from conventional methods for optimising chemical processes, which are essential for synthesising and improving prospective pharmacological compounds ([41]). These methods rely on well-founded ideas of organic chemistry, which include empirical rules and refined reaction conditions throughout decades of laboratory work. The optimised reaction conditions maximise desired product reaction yields, purity, and selectivity through exact control of variables, including temperature, pressure, and catalysts. Although the optimisation of chemical reactions that are expensive and ineffective is still an essential feature of pharmaceutical synthesis, the conventional approach has generated a significant number of possible medications ([42]). Luckily, ML applications in chemical reaction optimisation can efficiently solve the restrictions of conventional approaches. Therefore, machine learning applications overcome the limitations of conventional approaches, such as LabMate.ML ([43]) is shown in Figure 3.

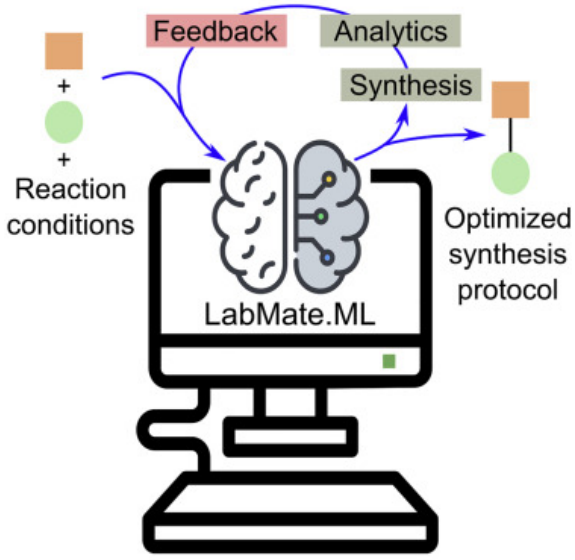


Figure 3: LabMate.ML Workflow for Automated Reaction Optimization

LabMate.ML simplifies the process of optimising 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 will collect various data points and enhance its predictive model. During iterations 11-20, LabMate.ML adopts an exploitative (greedy) strategy, prioritising selecting conditions that maximise the target value of the reaction output. The selection approach entails selecting reactions with the most minor 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, optimising 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 optimised 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 specialised gear. This method democratises the field of synthetic chemistry by combining chemical intuition with computational automation, resulting in improved efficiency and reduced resource usage ([43]).

Conventional, time-consuming, and expensive approaches have been proposed to be improved using machine learning techniques, such as LabMate.ML ([43]). Also, Aspuru-Guzik et al. effectively implemented automated chemical design utilising data-driven approaches ([44, 45, 46]) by using graph convolutional neural networks, especially specialised for molecular analysis. Building on graph convolution, Kearnes et al. ([47]) proposed one-shot learning as a tool to support drug discovery. The example shows how ML applied in chemical reaction optimisation can overcome the restrictions of conventional approaches for optimising chemical reactions.

1.6 Binding sites predictions

The binding site refers to a specific localised 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 ([48]). The most common approach to the identification of a binding site is the usage of structural information, such as dSPRINT ([49]) (Figure 4)

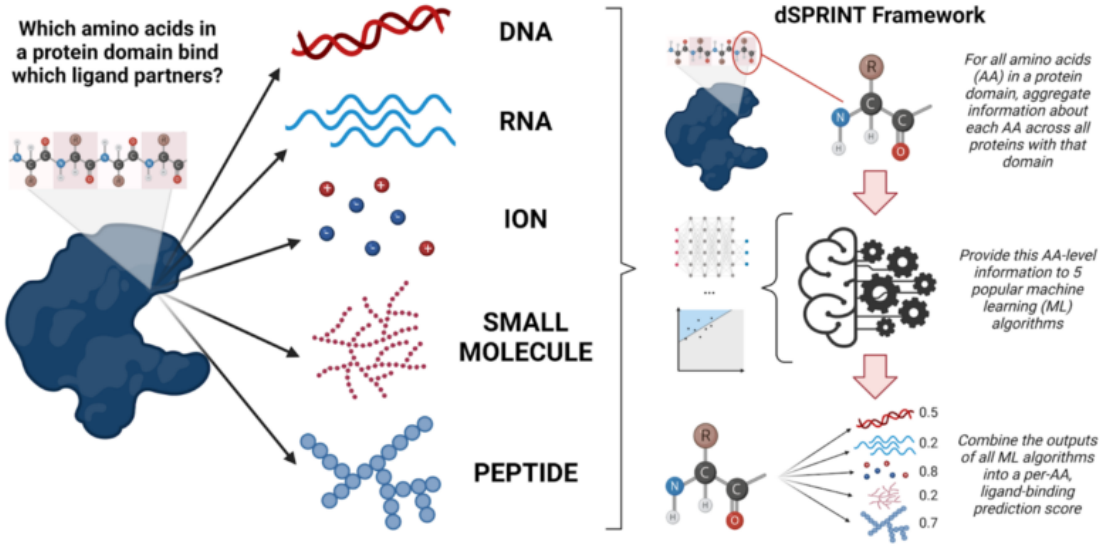


Figure 4: The representation of the dSPRINT (domain Sequence-based PRediction of INTeraction-sites) machine learning system for predicting protein-ligand interaction

Undergraduate student Anat Etzion-Fuchs from Mona Singh's lab at Princeton University leads the dSPRINT project, which forecasts the sites where ligands connect to protein domains using a range of traits, including amino acid patterns and physiochemical properties. Even for proteins without crystal structures, dSPRINT precisely forecasts interactions between domains and ligands by combining machine learning classifiers with a stacking approach. Cross-valuation validates the model's performance and reveals its capacity to forecast known interaction sites precisely. Furthermore, the forecasts of dSPRINT provide fresh insights into the interactions between proteins and ligands and expose new roles for proteins whose study has not been much investigated. This helps us to understand cellular processes and project protein functions better. One may easily review and investigate the dSPRINT architecture and its predictions ([49]).

Besides dSPRINT ([49]) (Figure 4), 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 ([50]). For instance, the FunFold methods use binding residues ([51, 52]) 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 ([53, 54]). P2Rank ([55]) is another instance of structural-based approaches 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 ([55]). 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 ([55]). At last, this accelerates the whole drug development process ([55]). The examples show that using ML for binding site predictions offers a good substitute for past methods' time-consuming and costly character.

1.7 Drug repurposing with ML

Drug repositioning is the technique of assigning indications for drugs meant for another use than their initial one. One of the most well-known instances, for instance, is sildenafil citrate, first recommended for angina and then developed into Viagra, a commonly used drug for erectile dysfunction ([56]). Many drugs have been moved throughout early medical history, including some ancient ones. The formal coining of medication repositioning can be found in a 2004 paper by Ashburn and Thor ([57]). Initially, this paper described the method used to investigate new ways to effectively administer present medications for yet-unknown disease indications ([57]). Drug repositioning studies the possible efficacy of existing medications and their application to treat diseases outside their current therapeutic reach. Among the advantages of drug repositioning over more traditional methods of drug development include less time and money needed for the latter. Furthermore, the medication employed has already been tested initially and has shown safety in human subjects, enabling the elimination of Phase I clinical studies. Drug repositioning shortens the cycle of pharmacological development ([58]).

Traditional ML in drug repositioning often involves utilising illness manifestations, drug targets, and genes as features. This is achieved by establishing a dichotomous network of links between features and medications, where similar disease targets are observed for similar drug actions ([58]). In the field of drug repurposing, various machine-learning methodologies have been employed, such as Linear regression and logistic regression ([59]), Support Vector Machines (SVMs) ([60]), Random Forest (RF) ([61]), and deep learning models ([62]). The published studies prove that utilising ML drug repurposing presents a potent approach to discovering novel therapeutic applications for existing pharmaceuticals, thus expediting the drug development procedure.

1.8 Multi-target drug predictions

In the last decade, most drug development research has concentrated on identifying compounds that operate on a singular target ([63]). Nevertheless, the selective affinity for a specific target consistently results in the instability of the cellular metabolic system, significantly impacting the regular physiological processes of cells and subsequently giving rise to detrimental effects. The selective affinity development of single-target medicines is subject to significant limitations ([64]). Conversely, certain intricate illnesses, including drug-induced liver damage (DILI), are controlled by numerous targets and frequently pose challenges for treatment with a medicine that targets only one of them ([65]). Therefore, the necessity of multi-target medications in treating complicated disorders has been established ([66]). For example, Lim et al. developed a model to design multi-target drug

prediction ([67]) (Figure 5).

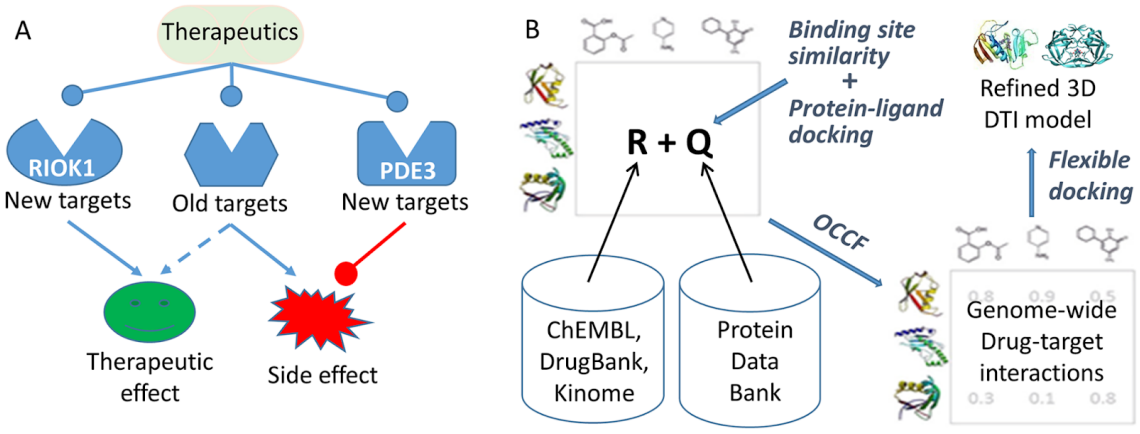


Figure 5: Integrated Strategy and Platform for Multi-Target Drug Screening.
(A) Illustration of a one-drug-multi-target-multi-indication strategy aimed at identifying drugs that enhance therapeutic effects while mitigating side effects, leveraging drug-off-target interactions. (B) Schema of 3D-REMAP, 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, 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 ([67]).

Chemical compounds classified as multi-target drugs combine the pharmacophores of numerous targets with different mechanisms of action into one molecule (Figure 5). These molecules may simultaneously interact with two or more molecular targets ([68]). A fresh strategy in multi-target drug development, polypharmacology—regulating numerous targets with one or more drugs ([69]) has attracted significant attention. This method mainly targets specific targets that particular medications can reasonably target ([70]). Thanks to substantial developments in the field of in silico pharmacology, several practical approaches for exploring multi-target medicines have emerged. Among these methods are pharmacophore modelling, network pharmacology, molecular docking, multi-omics-based system biology, Multi-target Quantitative Structure-Activity Relationship (mt-QSAR), machine learning, and perturbation model combined with machine learning (PLMT ([71])). Thus, it is expected that including different biological data and improving machine learning methods will help to optimise multi-target drug prediction models.

1.9 Biomarker predictions

ML-based biomarker identification and drug sensitivity predictive models are effective methods for enhancing clinical success rates, gaining deeper insights into pharmacological mechanisms, and identifying optimal treatments for individual patients ([72, 73])). Conducting late-stage clinical trials is a time-consuming and expensive process that spans several years and costs millions ([74])). Therefore, developing, validating, and utilising predictive models earlier would be most advantageous, using preclinical and early-stage clinical trial data. Translational biomarkers can be forecasted by employing ma-

chine learning techniques on preclinical datasets. Once the model and its accompanying biomarker have been validated using separate data sets from preclinical or clinical studies, they can be used to categorise patients, determine possible uses for medicine, and propose how the treatment works ([28])). For example, ML model used in COVID-19 to identify biomarkers ([75])) (Figure 6)

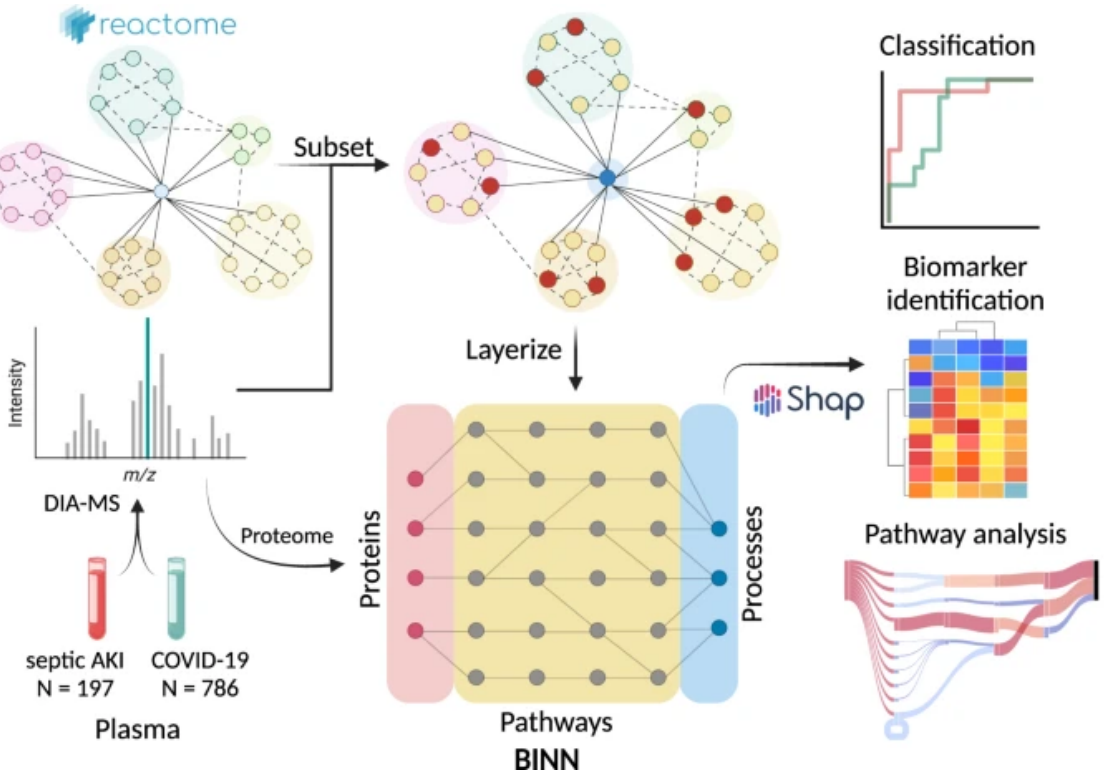


Figure 6: The representation of ML application in biomarker identification

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 organised into a sequential structure resembling a neural network. The protein amounts of each sample are utilised to train the corresponding BINNs to distinguish between two subphenotypes. Subsequently, the networks are analysed using SHAP, which provides feature importance values that can be used to identify biomarkers and do route analysis ([75])).

IntelliGenes, another ML application in biomarker identification, is a revolutionary ML pipeline for multi-genomics biomarker research with excellent accuracy for illness prediction ([76]). Using multi-genomic, clinical, and demographic data, this unique approach combines statistical methods with ML algorithms ([76]). At the heart of IntelliGenes is the Intelligent Gene (I-Gene) score, a new metric that assesses biomarkers' ability to predict complex features. IntelliGenes uses Recursive Feature Elimination, Random Forest, Support Vector Machine, XGBoost, k-Nearest Neighbours, Multi-Layer Perceptron, and soft and hard voting classifiers with Pearson correlation, Chi-square test, and ANOVA. IntelliGenes can find disease-associated biomarkers that traditional approaches miss with this combo. The method starts with AI/ML-ready Clinically Integrated Ge-

nomics and Transcriptomics (CIGT) data, including patient age, gender, race, ethnicity, diagnoses, and RNA-seq gene expression data. IntelliGenes ranks the top percentage of multi-genomic data-based profiles to predict diagnoses accurately. Also, IntelliGenes' personalised early diagnosis of common and unusual diseases is a significant benefit. Biomarkers' relevance in disease prediction is categorised by the novel I-Gene score. The I-Gene score measures biomarkers' weighted utility and biological system expression using SHAP and HHI. This directionality lets researchers establish whether biomarker overexpression or underexpression causes disease, revealing disease processes. IntelliGenes' versatility and adaptability allow it to be used on personal devices and high-performance computing platforms ([76]). This makes it an effective tool for research and clinical applications, enabling personalised interventions and novel therapy targets ([76]). The examples illustrate the great potential of machine learning applications in predicting biomarkers for drug research and development.

1.10 Computational pathology predictions

Computational pathology predictions are a breakthrough approach in pathology using advanced algorithms and machine-learning techniques to increase diagnosis accuracy and efficiency. These computer methods use vast histological data to produce reliable, consistent results above accepted manual interpretations. For instance, Lee et al. have demonstrated that using computational techniques to analyse the surrounding non-cancerous tissue in prostate cancer can yield important information on the chance of cancer returning ([77]). In oral malignancies and early-stage oestrogen receptor-positive breast tumours, Lu et al. showed a strong association between survival rates and characteristics defining the shape ([78]) and orientation of the nucleus ([79]). Such a combination of computational prediction in pathology reduces human error and accelerates the diagnosis process, thereby improving patient outcomes. The continuous evolution of computational pathology has significant power to change tailored treatment and enable the discovery of fresh biomarkers. The combination of technology and pathology (Figure 7) can transform healthcare by improving diagnosis accuracy, allowing targeted therapies, and raising the general quality of medical treatment.

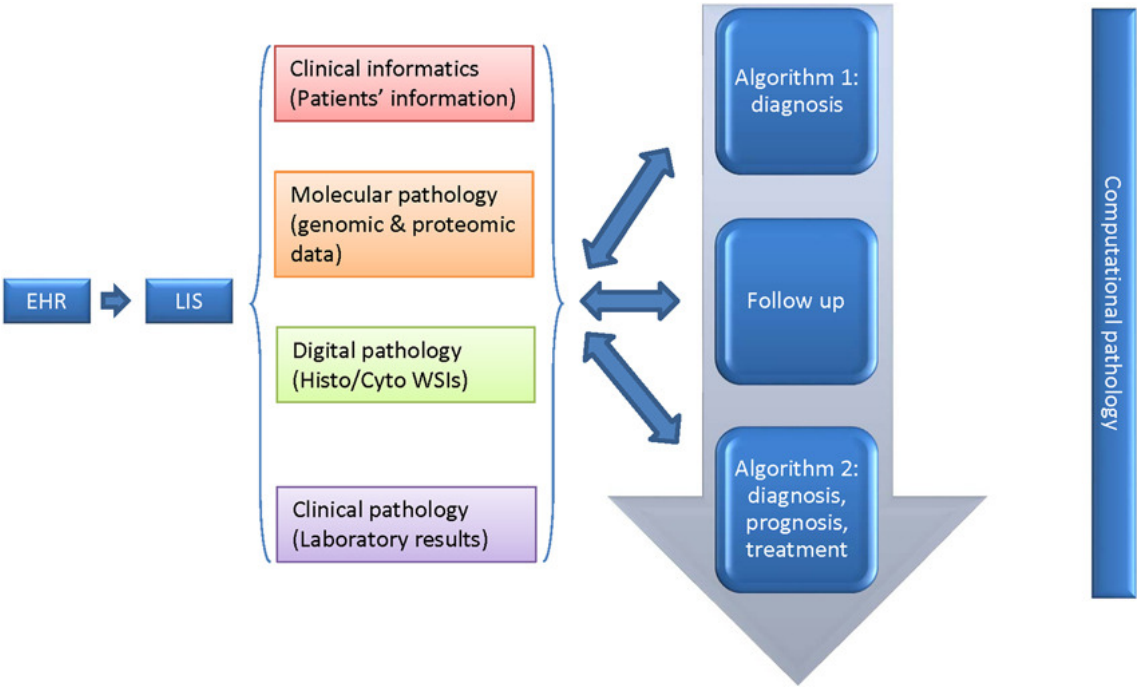


Figure 7: Medical System with Pathologist-Centered Approach.

The graphic shows a thorough, pathologist-centered medical system combining many data sources for improved diagnosis and patient care. Aggregated into the Laboratory Information System (LIS), clinical data from electronic health records (EHR), omics data from molecular pathology, whole slide images (WSIs) from digital pathology, and clinical laboratory results are included. This integration is the foundation of "Algorithm 1," a diagnostic tool synthesising several kinds of data to raise diagnostic accuracy. Updated disease-related data gathered during patient follow-up over time is merged with the original dataset to develop "Algorithm 2." Through constant improvement of diagnostic and prognostic capacities made possible by this iterative process, more exact and individualised patient treatment finally results ([80]).

1.11 Homology modelling/prediction of protein folding and stability of proteins upon mutations

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 ([81]). However, AlphaFold ([82]) initiates its computational exploration for the probable structure by utilising a template, a pre-existing structure for similar proteins. Alternatively, AlphaFold can utilise data on the biological evolution of amino-acid sequences within the protein family, which consists of proteins with similar activities and presumably possess identical folds ([82]). 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 ([83, 84]). Information of this nature can be derived from protein families' multiple sequence alignments (MSAs), obtained by analysing 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 ([85]).

1.12 Pharmacokinetics and Pharmacodynamics (PK/PD) predictions

Discovering and developing medications, regulatory review, and authorisation depend critically on pharmacokinetic (PK) and pharmacokinetic/pharmacodynamic (PK/PD) models. The design of clinical trials, data interpretation, and dose ([86]) computation depends critically on pharmacokinetic (PK) and pharmacokinetic/pharmacodynamic (PK/PD) models. Applying machine learning technology for PK and PK/PD prediction offers not only better results but also faster forecasts for this critical chore (Figure 8).

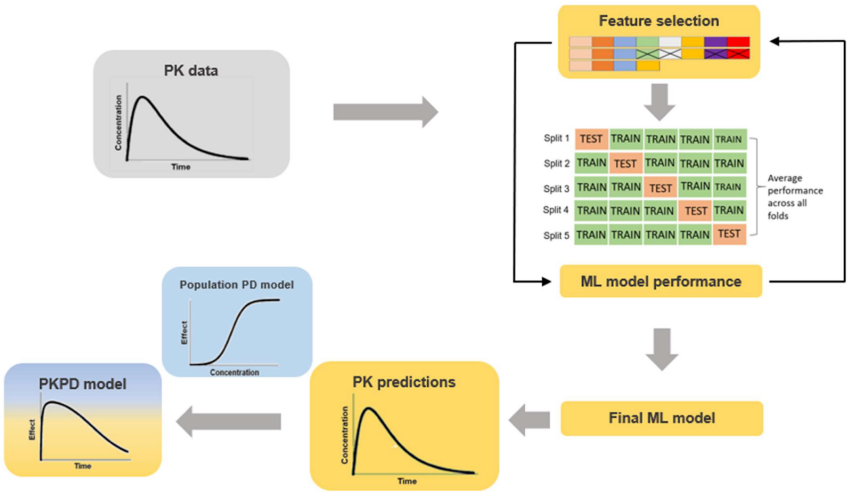


Figure 8: Proposed Workflow for Predicting Pharmacokinetics and Pharmacodynamics (PK/PD). .

This figure illustrates a sample process that combines pharmacometrics and machine learning to provide predictions about pharmacokinetics (PK) and pharmacodynamics (PD). The blue panels reflect components related to pharmacometrics, while the yellow panels indicate processes pertaining to machine learning. The framework established in this model utilises machine learning approaches to analyse and interpret pharmacometric data, hence enabling PK/PD prediction. By seamlessly integrating various methodologies, the model improves the precision and effectiveness of PK/PD prediction, assisting in optimising drug dosage and effectiveness ([87])

The emergence of advanced DL neural networks has resulted in the extensive utilisation of DL across various domains, including the field of medication discovery. Lately, there has been a surge of interest in utilising neural networks for PK and PK/PD modelling ([88]). In 2021, Liu et al. ([89]) employed the long short-term memory (LSTM) network to examine a hypothetical drug's PK/PD data using an indirect response model. They successfully demonstrated that the LSTM network effectively captured the temporal dependencies within the PK/PD relationship. Also, Liu et al. faced notable constraints when using the LSTM, such as its limited capacity to accurately predict dose levels beyond the training data and its inability to effectively represent sparse data ([89]). More recently, In 2022, Liu et al. ([90]) demonstrated that neural networks outperformed classical multinomial logistic regression in modelling the exposure-response connection. The artificial neural network (ANN) provided more accurate treatment effects forecasts than the regression technique, with less bias ([90, 91]). The successful utilisation of ML in PK and PK/PD prediction showcases the promising potential of ML to improve performance in drug discovery and development.

1.13 Drug-Drug Interaction Predictions

A drug-drug interaction (DDI) is the condition wherein another drug ingested or supplied concurrently modulates the way a drug is metabolised in the body (pharmacokinetics) or its effects on the body (pharmacodynamics) ([92]). Drug-drug interactions (DDIs) can cause products to be removed from the market; for example, astemizole, a medication used to treat allergy symptoms, was taken off the market because it can cause the QT interval to become longer and lead to irregular heart rhythms when combined with substances that inhibit cytochrome P450 3A4 (CYP3A4), such grapefruit juice and erythromycin ([93]). Taken off the market due to a sluggish heart rate and muscle breakdown when used with some cardiovascular drugs such as beta-blockers or statins, Mibefradil was used to treat excessive blood pressure and chronic chest discomfort ([94]). Thus, after time-consuming and expensive drug research and development phases, it is imperative to effectively ascertain drug-drug interactions to prevent withdrawals by employing promising ML models subsequently. For example, a deep learning framework called Deepdrug has investigated drug-drug interaction in Figure 9 ([95]).

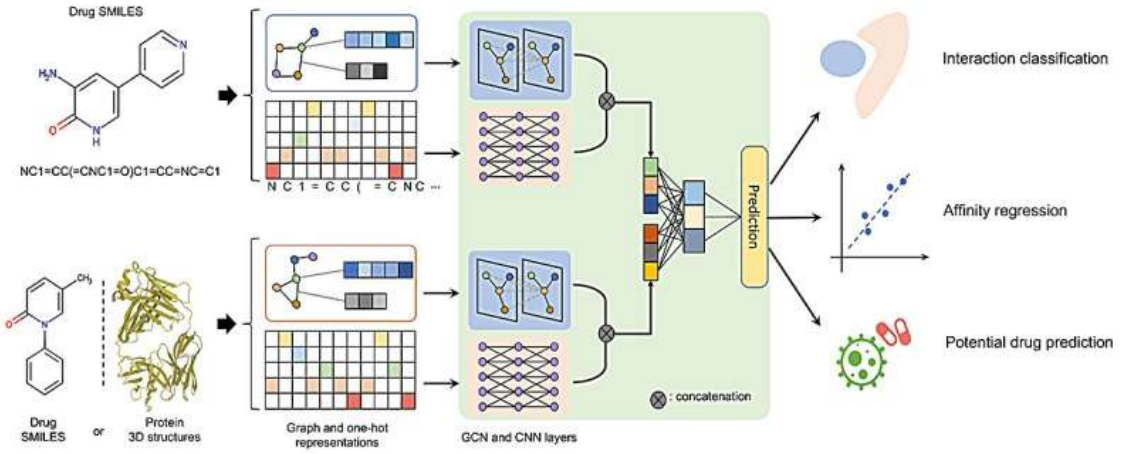


Figure 9: A deep learning framework is developed for predicting drug interactions and affinity.

This 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 ([95]).

Besides Deepdrug Figure 9 ([95]), several machine-learning methods have been created to forecast drug-drug interactions (DDIs) to address the shortage of known DDI combinations ([96, 97, 98]). Prior research has yielded numerous models for predicting the occurrence or non-occurrence of drug-drug interactions (DDIs). These models have successfully identified DDI combinations that result in adverse effects and categorised

the various forms of DDIs by utilising publicly available databases. However, the models showed limited success in identifying DDI combinations ([99]).

The models used in drug-drug interaction (DDI) predictions exhibit evident constraints. These limitations often stem from incomplete data, variability in patient responses, and the complexity of accurately modelling the factors influencing DDIs ([99]). Also, most models have solely offered a fundamental forecast for the presence or categorization of DIs ([100]). These models do not help make complex therapeutic decisions, such as accurately adjusting the dosage or selecting alternative drugs. Physicians and chemists require predictions on the fold change of PK parameters. However, thus far, no models have successfully forecasted this. Also, there is currently no available dataset that has been methodically developed to contain true-negative instances exclusively. For example, the DDI DB, such as DrugBank, is commonly utilised for predicting drug-drug interactions (DDIs). It provides information indicating the presence of a DDI between drugs A and B but does not include information indicating the absence of a DDI. Therefore, researchers have unavoidably chosen random sets of medication combinations under the assumption that there were no interactions ([101, 102]). Indeed, the lack of proof does not serve as evidence for the absence of something. Utilising a model lacking in quality input poses challenges in acquiring dependable results. When the negative set is randomly generated, it becomes challenging to pinpoint the specific reason for unanticipated problematic output ([103]). As a result, there is a continuous need for more sophisticated and robust ML models to better account for these variables and improve the reliability of DDI predictions.

1.14 Clinical Trial Simulations with ML

Clinical research includes both interventional and observational studies that utilise volunteers to assess the efficacy and safety of novel drugs, treatments, or medical devices ([104]). The design of clinical trials is a crucial element in interventional studies, as it enables a thorough evaluation of the effectiveness and safety of therapies ([105]). ClinicalTrials.gov houses more than 400,000 records of clinical studies conducted in 220 countries. The sponsor or primary investigator of the clinical research is responsible for registering the studies on the website. In 2022, 14.5% of interventional trials listed in the clinicaltrials.gov registry ended prematurely ([106, 107]).

The usual methods to replicate clinical trials primarily relied on deterministic models and simple probabilistic approaches to replicate trial circumstances and evaluate prospective outcomes ([108]). More complex models that consider patient demographics, illness development, and therapeutic effects can be produced as computer capability develops (Figure 10). By predicting anticipated difficulties and improving trial procedures, clinical trial simulations have helped to raise the ethical standards and efficiency of trial design. These approaches created the basis for the present advanced simulations, which today incorporate sophisticated algorithms and machine learning techniques to produce even more exact and dynamic forecasts ([109]), even if they have some constraints.

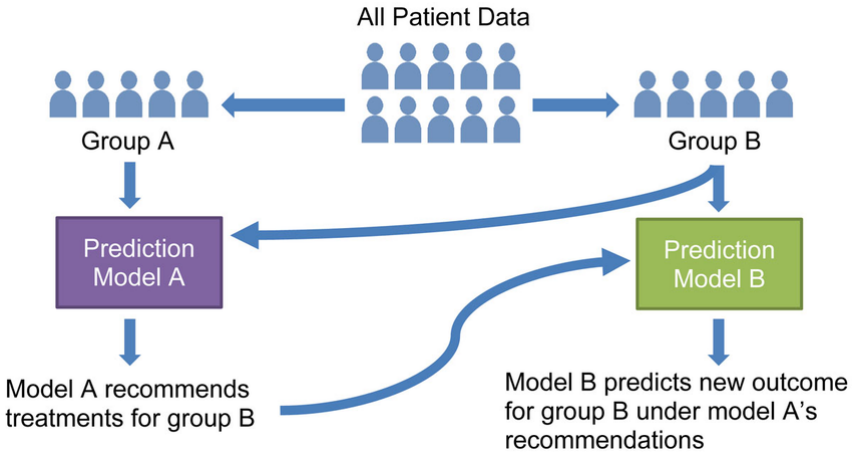


Figure 10: The representation of simulation of clinical trials. Clinical trial simulation procedure.

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 ([110]).

Multiple studies in the literature have concentrated on utilising machine learning algorithms to forecast the success of clinical trials ([111, 112, 113]). These algorithms are based on structured and unstructured information obtained from clinical studies. These investigations aim to predict the premature termination of clinical trials by utilising trial characteristic data in conjunction with unstructured data. Follett et al. utilised both structured and unstructured data to forecast the discontinuation of clinical trials ([114]). The study employed text mining techniques to extract vector features from the studies' description fields and the studies' fundamental aspects. A dataset of 130,000 investigations was compiled by selecting completed and terminated trials using accessible information before the study commenced. Subsequently, they employed machine learning methodologies to forecast results ([107]). Also, The study done by Elkin et al. ([115]) employed machine learning techniques to predict the termination of clinical trials based on data taken from clinicalTrials.gov ([106]). The study produced a dataset consisting of 68,999 studies. A total of 640 features were created by employing document embedding, keyword features, and trial characteristics. Their study aimed to comprehensively understand clinical trial termination and produce accurate prediction outcomes ([107]). Machine learning (ML) in clinical trial simulations demonstrates its potential to enhance performance by reducing the time required.

1.15 Drug delivery simulations with ML

In pharmacological research and development, applying machine learning (ML) in drug delivery simulations marks a radical change ([116, 117]). Often, depending on empirical data and crude computational models, conventional drug delivery systems can be time-consuming and less exact. Still, machine learning offers a vital substitute by using large databases to predict drug behaviour, streamline delivery systems, and raise treatment outcome accuracy ([118]). Using cutting-edge algorithms and data-driven insights, ML-driven drug delivery simulations faithfully replicate complex biological systems, identify

the most efficient delivery paths, and project tailored patient reactions. This approach not only accelerates the development of new pharmaceuticals but also provides opportunities for tailored medicine, ensuring that treatments are adjusted to fit every patient's particular requirements and produce the best possible effectiveness and the fewest possible side effects. The advent of machine learning (ML) in medication delivery models has the power to revolutionise the whole industry. Increasing speed, efficiency, and accuracy above past capabilities ([119, 120]) can improve drug development (Figure 11).

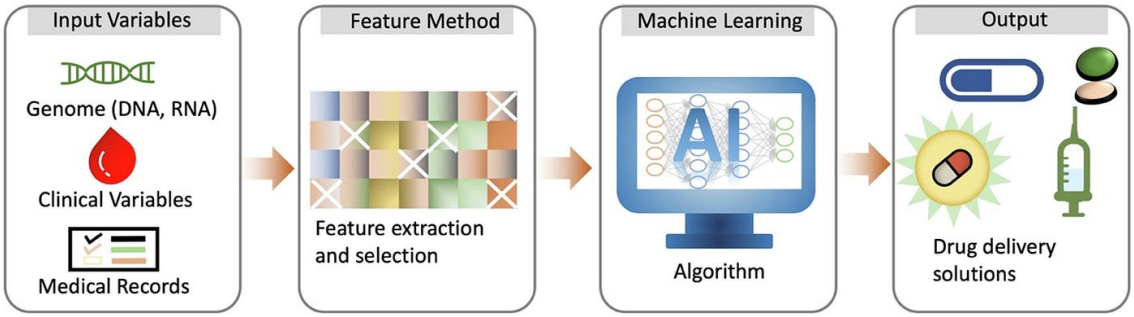


Figure 11: The using machine learning methodologies in drug delivery simulation to improve pharmaceutical research and development.

The pharmaceutical industry is increasingly adopting AI techniques, including machine learning, to enhance the effectiveness of drugs in terms of their bioavailability, stability, and ability to target specific organs. Computational pharmaceutics transforms medication delivery paradigms by utilising big data and multiscale modelling techniques. Machine learning allows for identifying relevant characteristics, the discovery of unexpected patterns, and immediately adjusting to pathogen evolution, hence allowing effective treatment development, selection, and dosage optimisation. Machine learning integrated into clinical devices can significantly transform clinical decision-making in anti-infective drug delivery strategies, thanks to its impressive speed and practicality ([121]).

One of the main advantages of applying artificial intelligence in medical delivery systems is its capacity to handle and evaluate large amounts of data effectively. Artificial intelligence algorithms can use large amounts of data to identify trends in medication formulations that might not be readily apparent using more traditional approaches ([122, 123]). Artificial intelligence methods can examine data-driven sources such as chemical structures, physicochemical properties, and biological activity data to expose hidden relationships. Furthermore, enabling the forecast of medication durability in different storage conditions, the solubility of pharmaceuticals in various formulations, and the release of kinetics from other delivery systems is the ability of machine learning models to grasp complex connections and non-linear interdependencies ([124, 125]). Using pattern and trend analysis, AI can thus enhance medicine delivery systems, influencing ([117, 126, 123]) greater effectiveness, safety, stability, solubility, and release kinetics.

1.16 Challenges and Limitations of Machine Learning in Drug Discovery

Although machine learning offers excellent possibilities to change drug discovery processes, it is not free from challenges and restrictions. The accessibility and quality of data create a significant challenge since the effectiveness of machine learning models

depends much on large, diversified, carefully kept datasets ([127]). Furthermore, a considerable challenge of machine learning models is their comprehensibility since complex algorithms usually produce predictions without explaining underlying mechanisms. Furthermore, the opaque character of many machine learning models could hinder regulatory approval and acceptance inside the pharmaceutical company ([128]). Moreover, the ability of machine learning models to be applied to various datasets and experimental environments is still tricky since models trained on particular datasets could not attain optimal performance on fresh, undiscovered data ([129]). Dealing with these challenges calls for cooperation between several fields, advances data organisation and standardising, and generates interpretable and transferable machine learning models specially tailored for the complex character of drug development.

2 Emerging Trends in Drug Discovery

Over the last ten years, there have been significant advancements in the field of cancer biology. These include major breakthroughs in genomics technology, which have allowed scientists to obtain DNA sequences of thousands of tumours. Additionally, researchers have been able to analyse the gene expression profiles of tumours at the individual cell level. Furthermore, the functional properties of genes have been studied through comprehensive CRISPR screens of the entire genome, leading to the ident. Regrettably, many therapeutic targets arising from these advancements have been classified as "undruggable" based on traditional criteria ([130]). These targets often include extensive protein-protein interaction surfaces that do not have readily identifiable pockets for binding ligands. However, the drug discovery community has shown a growing commitment to the challenge of developing drugs for targets that were previously considered impossible to drug. As a result, researchers have focused on developing new methods to target previously considered "undruggable" targets, such as KRAS ([130]). The targeting of KRAS with drugs aims to analyse the variables that led to the development of innovative approaches to address this challenging and difficult-to-drug target ([130]). Another successful example of targeting "undruggable" targets is PROTAC ([131]). PROTAC can degrade the protein of interest using the ubiquitin mechanism in the cell ([131]). The last successful example is that Prodrugs have been used in the targeting of "undruggable" targets ([132]).

2.1 Drugging KRAS

KRAS is an oncogene that is commonly mutated in human cancer, with mutations occurring in around 30% of all tumours ([130]). 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 centred on posttranslational 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 ([133, 134, 135]).

2.2 Prodrugs

In cancer treatment, prodrugs like PRIMA-1 and its derivative PRIMA-1MET present a viable method for attacking "undruggable" targets such 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 re-establishing an oxidative environment inside tumour 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 stabilise p53, therefore keeping its tumour-suppressing properties. These creative approaches draw attention to the possibilities for efficiently modifying difficult cancer targets by the use of prodrugs and tailored compounds ([132]).

2.3 Drugging "Undruggable" Cancer Targets

Although $KRAS^{G12C}$ has been successful, numerous recurring oncoproteins and tumour suppressors, including MYC, β-catenin, and p53, have not been targeted with drugs. Additionally, resistance to single-drug treatments remains a significant obstacle. Recent technological advancements and novel strategies have started to tackle these obstacles, converting previously difficult-to-treat targets into ones that can be targeted with drugs. Some examples of these advancements include kinase inhibitors, antiapoptotic BCL2 proteins, and transcription factors. Innovations have dramatically influenced the growth in this field in structure-based design, chemically induced protein degradation, and targeted covalent inhibition ([130]).

2.4 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 ([136]). 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 ([137]. 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 ([130]).

2.5 Proximity Induction as a Groundbreaking New Concept in Drug Discovery

Targeting poorly ligandable proteins like zinc finger transcription factors, proximity induction in drug discovery—best shown by IMiDs and PROTACs—uses tiny compounds to cause protein breakdown. By means of protein-protein interactions, sometimes known as "molecular glue," IMiDs recruit CRBN to break down neo-substrates such as IKZF1/3, hence extending druggable targets. Pioneered by Crews and Bradner, PROTACs degrade proteins using E3 ligases, hence advancing into clinical trials for AR and ER. Although selective degradation and wide target space have benefits, practical problems include cellular penetration and target ligand availability. Glue molecules provide modest, empirical solutions, as shown by IMiDs, therefore extending degradation beyond CRBN. Discovered by Ebert's team, BI-3802 and CR8 both represent small molecule degraders aiming against BCL6 and CDK, respectively. These compounds enhance already existing protein-protein interactions, hence promoting breakdown. The classic molecular glue, cyclosporin, hooks cyclophilin to block calcineurin, hence modulating IL2 synthesis. Companies such as Revolution Medicines and WarpDrive Bio apply similar ideas for new therapies like KRASG12C inhibitors. By extending the range of druggable targets, these strategies should transform drug development ([138, 139, 140]).

2.6 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 ribocil, has 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 ([141, 142]).

2.7 Overcoming Drug Resistance

To tackle drug resistance in targeted treatments, it is necessary to comprehend both genetic and nongenetic causes. These mechanisms include mutations in the gene targeted by the medication and the adaptive reprogramming of signalling pathways. Novel pharmaceuticals, such as combination therapy and bivalent compounds, have been created to counteract resistance mutations in different cancer-causing sites, prolonging treatment effectiveness. Nevertheless, resistance to specific protein degraders such as PROTACs and molecular glues may develop. This would need the implementation of techniques to counteract resistance caused by mutations. Furthermore, resistance to drugs that do not involve mutations, such as adaptive reprogramming and cell plasticity, poses difficulties but also provides potential for rational medication combinations and treatments that target epigenetic changes to avoid or overcome resistance. Small compounds that target transcription factors are a valuable addition to the range of tools available to combat drug resistance in cancer treatment ([143, 144, 140, 145]).

2.8 Summary of Emerging Trends in Drug Discovery

Mainly by the use of genomics technology, single-cell gene expression analysis, and thorough CRISpen screens, significant advances in cancer biology over the past ten years have revolutionised medication development. Many therapeutic targets these developments found were first judged "undruggable" because of sizeable protein-protein interaction surfaces devoid of clear ligand-binding pockets. Still, the scientific community has embraced creating strategies to target these challenging proteins. Among the notable achievements are creating medicines aimed at KRASG12C, covalent bonding techniques to take advantage of latent binding sites, and PROTAC technology—which uses the ubiquitin-proteasome system to break down target proteins.

These days, new trends in drug development centre on creative lead-finding tools and fresh drug modalities. For typically "undruggable" proteins, proximity-based platforms, including chemoproteomic techniques, have evolved as valuable tools for locating covalent ligands and possible binding sites. Furthermore, the idea of proximity induction—as shown by IMiDs and PROTACs—has created fresh paths for inducing protein breakdown and thereby targeting less ligandable proteins. Targeting RNA also presents fresh opportunities for challenging-to-treat cancer targets since it seems a complementary approach to protein targeting. These developments are transforming the field of drug discovery and broadening the range of druggable targets when combined with attempts to overcome drug resistance with fresh drugs and combination therapies.

3 Conclusion

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.

References

[1] Hao Zhu. Big data and artificial intelligence modeling for drug discovery. *Annual review of pharmacology and toxicology*, 60(1):573–589, 2020.

[2] Lauv Patel, Tripti Shukla, Xiuzhen Huang, David W Ussery, and Shanzhi Wang. Machine learning methods in drug discovery. *Molecules*, 25(22):5277, 2020.

[3] Elina Petrova. Innovation in the pharmaceutical industry: The process of drug discovery and development. In *Innovation and Marketing in the Pharmaceutical Industry: Emerging Practices, Research, and Policies*, pages 19–81. Springer, 2013.

[4] Jiashun Mao, Javed Akhtar, Xiao Zhang, Liang Sun, Shenghui Guan, Xinyu Li, Guangming Chen, Jiaxin Liu, Hyeon-Nae Jeon, Min Sung Kim, et al. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *Iscience*, 24(9), 2021.

[5] Javier Vázquez, Manel López, Enric Gibert, Enric Herrero, and F Javier Luque. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules*, 25(20):4723, 2020.

[6] Periyasamy Natarajan Shiammala, Navaneetha Krishna Bose Duraimutharasan, Baskaralingam Vaseeharan, Abdulaziz S Alothaim, Esam S Al-Malki, Babu Snekaa, Sher Zaman Safi, Sanjeev Kumar Singh, Devadasan Velmurugan, and Chandrabose Selvaraj. Exploring the artificial intelligence and machine learning models in the context of drug design difficulties and future potential for the pharmaceutical sectors. *Methods*, 2023.

[7] Sonal Priya, Garima Tripathi, Dev Bukhsh Singh, Priyanka Jain, and Abhijeet Kumar. Machine learning approaches and their applications in drug discovery and design. *Chemical Biology & Drug Design*, 100(1):136–153, 2022.

[8] Song Xia, Eric Chen, and Yingkai Zhang. Integrated molecular modeling and machine learning for drug design. *Journal of chemical theory and computation*, 19(21):7478–7495, 2023.

[9] Hugo Jair Escalante. Automated machine learning—a brief review at the end of the early years. *Automated Design of Machine Learning and Search Algorithms*, pages 11–28, 2021.

[10] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[11] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[12] Kristy A Carpenter and Xudong Huang. Machine learning-based virtual screening and its applications to alzheimer's drug discovery: a review. *Current pharmaceutical design*, 24(28):3347–3358, 2018.

[13] Thomas J Hwang, Daniel Carpenter, Julie C Lauffenburger, Bo Wang, Jessica M Franklin, and Aaron S Kesselheim. Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA internal medicine*, 176(12):1826–1833, 2016.

[14] Claudio N Cavasotto and Valeria Scardino. Machine learning toxicity prediction: latest advances by toxicity end point. *ACS omega*, 7(51):47536–47546, 2022.

[15] Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 2022.

[16] Melvin E Andersen and Daniel Krewski. Toxicity testing in the 21st century: bringing the vision to life. *Toxicological sciences*, 107(2):324–330, 2009.

[17] Yunyi Wu and Guanyu Wang. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *International journal of molecular sciences*, 19(8):2358, 2018.

[18] VK Agrawal and PV Khadikar. Qsar prediction of toxicity of nitrobenzenes. *Bioorganic & medicinal chemistry*, 9(11):3035–3040, 2001.

[19] Anjali Setiya, Vinod Jani, Uddhavesh Sonavane, and Rajendra Joshi. Moltoxpred: small molecule toxicity prediction using machine learning approach. *RSC advances*, 14(6):4201–4220, 2024.

[20] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

[21] Bhanushee Sharma, Vijil Chenthamarakshan, Amit Dhurandhar, Shiranee Pereira, James A Hendler, Jonathan S Dordick, and Payel Das. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Scientific Reports*, 13(1):4908, 2023.

[22] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in bioinformatics*, 22(1):247–269, 2021.

[23] Frank W Pun, Ivan V Ozerov, and Alex Zhavoronkov. Ai-powered therapeutic target discovery. *Trends in pharmacological sciences*, 2023.

[24] Gerhard Wolber, Thomas Seidel, Fabian Bendix, and Thierry Langer. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug discovery today*, 13(1-2):23–29, 2008.

[25] James H Nettles, Jeremy L Jenkins, Andreas Bender, Zhan Deng, John W Davies, and Meir Glick. Bridging chemical and biological space:"target fishing" using 2d and 3d molecular descriptors. *Journal of medicinal chemistry*, 49(23):6802–6810, 2006.

[26] Yu-Chen Lo, Silvia Senese, Robert Damoiseaux, and Jorge Z Torres. 3d chemical similarity networks for structure-based target prediction and scaffold hopping. *ACS chemical biology*, 11(8):2244–2253, 2016.

[27] Aeri Lee, Kyoungyeul Lee, and Dongsup Kim. Using reverse docking for target identification and its applications for drug discovery. *Expert opinion on drug discovery*, 11(7):707–715, 2016.

[28] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[29] Jintae Kim, Sera Park, Dongbo Min, and Wankyu Kim. Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, 22(18):9983, 2021.

[30] Romano T Kroemer. Structure-based drug design: docking and scoring. *Current protein and peptide science*, 8(4):312–328, 2007.

[31] Shuyan Li, Lili Xi, Chengqi Wang, Jiazhong Li, Beilei Lei, Huanxiang Liu, and Xiaojun Yao. A novel method for protein-ligand binding affinity prediction and the related descriptors exploration. *Journal of computational chemistry*, 30(6):900–909, 2009.

[32] DiMasi Ja. The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 2002.

[33] Daniel K Gehlhaar, Gennady M Verkhivker, Paul A Rejto, Christopher J Sherman, David R Fogel, Lawrence J Fogel, and Stephan T Freer. Molecular recognition of the inhibitor ag-1343 by hiv-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & biology*, 2(5):317–324, 1995.

[34] Hans FG Velec, Holger Gohlke, and Gerhard Klebe. Drugscorecsd knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of medicinal chemistry*, 48(20):6296–6303, 2005.

[35] Todd JA Ewing, Shingo Makino, A Geoffrey Skillman, and Irwin D Kuntz. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design*, 15:411–428, 2001.

[36] Sangmin Seo, Jonghwan Choi, Sanghyun Park, and Jaegyoon Ahn. Binding affinity prediction for protein–ligand complex using deep attention mechanism based on intermolecular interactions. *BMC bioinformatics*, 22:1–15, 2021.

[37] Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[38] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

[39] Hongjian Li, Kwong-Sak Leung, Man-Hon Wong, and Pedro J Ballester. Improving autodock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular informatics*, 34(2-3):115–126, 2015.

[40] Zhan Deng, Claudio Chuaqui, and Juswinder Singh. Structural interaction fingerprint (sift): a novel method for analyzing three-dimensional protein- ligand binding interactions. *Journal of medicinal chemistry*, 47(2):337–344, 2004.

[41] KC Nicolaou. Advancing the drug discovery and development process. *Angewandte Chemie*, 126(35):9280–9292, 2014.

[42] Zhenpeng Zhou, Xiaocheng Li, and Richard N Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS central science*, 3(12):1337–1344, 2017.

[43] Daniel Reker, Emily A Hoyt, Gonçalo JL Bernardes, and Tiago Rodrigues. Adaptive optimization of chemical reactions with minimal experimental information. *Cell Reports Physical Science*, 1(11), 2020.

[44] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[45] José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. In *International conference on machine learning*, pages 1470–1479. PMLR, 2017.

[46] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[47] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30:595–608, 2016.

[48] Janez Konc. Binding site comparisons for target-centered drug discovery. *Expert Opinion on Drug Discovery*, 14(5):445–454, 2019.

[49] Anat Etzion-Fuchs, David A Todd, and Mona Singh. dsprint: predicting dna, rna, ion, peptide and small molecule interaction sites within protein domains. *Nucleic acids research*, 49(13):e78–e78, 2021.

[50] Charles A Santana, Sandro C Izidoro, Raquel C de Melo-Minardi, Jonathan D Tyzack, Antonio JM Ribeiro, Douglas EV Pires, Janet M Thornton, and Sabrina de A. Silveira. Grasp-web: a machine learning strategy to predict binding sites based on residue neighborhood graphs. *Nucleic Acids Research*, 50(W1):W392–W397, 2022.

[51] Daniel B Roche, Stuart J Tetchner, and Liam J McGuffin. Funfold: an improved automated method for the prediction of ligand binding residues using 3d models of proteins. *BMC bioinformatics*, 12:1–20, 2011.

[52] Daniel B Roche, Maria T Buenavista, and Liam J McGuffin. The funfold2 server for the prediction of protein–ligand interactions. *Nucleic acids research*, 41(W1):W303–W307, 2013.

[53] Jianyi Yang, Ambrish Roy, and Yang Zhang. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, 29(20):2588–2595, 2013.

[54] Qi Wu, Zhenling Peng, Yang Zhang, and Jianyi Yang. Coach-d: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic acids research*, 46(W1):W438–W442, 2018.

[55] Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10:1–12, 2018.

[56] Manli Wang, Ruiyuan Cao, Leike Zhang, Xinglou Yang, Jia Liu, Mingyue Xu, Zhengli Shi, Zhihong Hu, Wu Zhong, and Gengfu Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-ncov) in vitro. *Cell research*, 30(3):269–271, 2020.

[57] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683, 2004.

[58] Lijun Cai, Jiaxin Chu, Junlin Xu, Yajie Meng, Changcheng Lu, Xianfang Tang, Guanfang Wang, Geng Tian, and Jialiang Yang. Machine learning for drug repositioning: Recent advances and challenges. *Current Research in Chemical Biology*, page 100042, 2023.

[59] Huiyuan Chen and Jing Li. A flexible and robust multi-source learning algorithm for drug repositioning. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 510–515, 2017.

[60] Zaynab Mousavian, Sahand Khakabimamaghani, Kaveh Kavousi, and Ali Masoudi-Nejad. Drug–target interaction prediction from pssm based evolutionary information. *Journal of pharmacological and toxicological methods*, 78:42–51, 2016.

[61] Han Shi, Simin Liu, Junqi Chen, Xuan Li, Qin Ma, and Bin Yu. Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics*, 111(6):1839–1852, 2019.

[62] Hui Liu, Wenhao Zhang, Yinglong Song, Lei Deng, and Shuigeng Zhou. Hnet-dnn: inferring new drug–disease associations with deep neural network based on heterogeneous network features. *Journal of Chemical Information and Modeling*, 60(4):2367–2376, 2020.

[63] Richard Morphy, Corinne Kay, and Zoran Rankovic. From magic bullets to designed multiple ligands. *Drug discovery today*, 9(15):641–651, 2004.

[64] Richard B Rothman, Michael H Baumann, Jason E Savage, Laura Rauser, Ace McBride, Sandra J Hufeisen, and Bryan L Roth. Evidence for possible involvement of 5-ht2b receptors in the cardiac valvulopathy associated with fenfluramine and other serotonergic medications. *Circulation*, 102(23):2836–2841, 2000.

[65] Mark Real, Michele S Barnhill, Cory Higley, Jessica Rosenberg, and James H Lewis. Drug-induced liver injury: highlights of the recent literature. *Drug safety*, 42:365–387, 2019.

[66] Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008.

[67] Hansaim Lim, Di He, Yue Qiu, Patrycja Krawczuk, Xiaoru Sun, and Lei Xie. Rational discovery of dual-indication multi-target pde/kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology. *PLoS computational biology*, 15(6):e1006619, 2019.

[68] Richard Morphy and Zoran Rankovic. Multi-target drugs: Strategies and challenges for medicinal chemists. In *The Practice of Medicinal Chemistry*, pages 549–571. Elsevier, 2008.

[69] Antonio Lavecchia and Carmen Cerchia. In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discovery Today*, 21(2):288–298, 2016.

[70] Andrew Anighoro, Jurgen Bajorath, and Giulio Rastelli. Polypharmacology: challenges and opportunities in drug discovery: miniperspective. *Journal of medicinal chemistry*, 57(19):7874–7887, 2014.

[71] Kaiyang Liu, Xi Chen, Yue Ren, Chaoqun Liu, Tianyi Lv, Yanling Zhang, et al. Multi-target-based polypharmacology prediction (mtpp): an approach using virtual screening and machine learning for multi-target drug discovery. *Chemico-Biological Interactions*, 368:110239, 2022.

[72] Bin Li, Hyunjin Shin, Georgy Gulbekyan, Olga Pustovalova, Yuri Nikolsky, Andrew Hope, Marina Bessarabova, Matthew Schu, Elona Kolpakova-Hart, David Merberg, et al. Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. *PLoS one*, 10(6):e0130700, 2015.

[73] Virginia B Kraus. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nature Reviews Rheumatology*, 14(6):354–362, 2018.

[74] Jim Weatherall, Faisal M Khan, Mishal Patel, Richard Dearden, Khader Shameer, Glynn Dennis, Gabriela Feldberg, Thomas White, and Sajan Khosla. Clinical trials, real-world evidence, and digital medicine. In *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, pages 191–215. Elsevier, 2021.

[75] Erik Hartman, Aaron M Scott, Christofer Karlsson, Tirthankar Mohanty, Suvi T Vaara, Adam Linder, Lars Malmström, and Johan Malmström. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nature Communications*, 14(1):5359, 2023.

[76] William DeGroat, Dinesh Mendhe, Atharva Bhusari, Habiba Abdelhalim, Saman Zeeshan, and Zeeshan Ahmed. Intelligenes: a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. *Bioinformatics*, 39(12):btad755, 2023.

[77] George Lee, Robert W Veltri, Guangjing Zhu, Sahirzeeshan Ali, Jonathan I Epstein, and Anant Madabhushi. Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. *European urology focus*, 3(4-5):457–466, 2017.

[78] Cheng Lu, James S Lewis Jr, William D Dupont, W Dale Plummer Jr, Andrew Janowczyk, and Anant Madabhushi. An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk

stratify patients for disease-specific survival. *Modern Pathology*, 30(12):1655–1665, 2017.

[79] Cheng Lu, David Romo-Bucheli, Xiangxue Wang, Andrew Janowczyk, Shridar Ganesan, Hannah Gilmore, David Rimm, and Anant Madabhushi. Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers. *Laboratory investigation*, 98(11):1438–1448, 2018.

[80] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021.

[81] Seth D Axen, Xi-Ping Huang, Elena L Cáceres, Leo Gendelev, Bryan L Roth, and Michael J Keiser. A simple representation of three-dimensional molecular structure. *Journal of medicinal chemistry*, 60(17):7393–7409, 2017.

[82] Alessia David, Suhail Islam, Evgeny Tankhilevich, and Michael JE Sternberg. The alphafold database of protein structures: a biologist's guide. *Journal of molecular biology*, 434(2):167336, 2022.

[83] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[84] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.

[85] Armita Nourmohammad, Michael Pun, and Gian Marco Visani. Machine-learning model reveals protein-folding physics. *Physics*, 15:183, 2022.

[86] Bernd Meibohm and Hartmut Derendorf. Basic concepts of pharmacokinetic/pharmacodynamic (pk/pd) modelling. *International journal of clinical pharmacology and therapeutics*, 35(10):401–413, 1997.

[87] Lina Keutzer, Huifang You, Ali Farnoud, Joakim Nyberg, Sebastian G Wicha, Gareth Maher-Edwards, Georgios Vlasakakis, Gita Khalili Moghaddam, Elin M Svensson, Michael P Menden, et al. Machine learning and pharmacometrics for prediction of pharmacokinetic data: differences, similarities and challenges illustrated with rifampicin. *Pharmaceutics*, 14(8):1530, 2022.

[88] James Lu, Brendan Bender, Jin Y Jin, and Yuanfang Guan. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. *Nature machine intelligence*, 3(8):696–704, 2021.

[89] Xiangyu Liu, Chao Liu, Ruihao Huang, Hao Zhu, Qi Liu, Sunanda Mitra, and Yaning Wang. Long short-term memory recurrent neural network for pharmacokinetic-pharmacodynamic modeling. *International journal of clinical pharmacology and therapeutics*, 59(2):138, 2021.

[90] Chao Liu, Yuan Xu, Qi Liu, Hao Zhu, and Yaning Wang. Application of machine learning based methods in exposure–response analysis. *Journal of Pharmacokinetics and Pharmacodynamics*, 49(4):401–410, 2022.

[91] Albert Tang. Machine learning for pharmacokinetic/pharmacodynamic modeling. *Journal of Pharmaceutical Sciences*, 112(5):1460–1475, 2023.

[92] Rakshit S Tanna. *Assessing the pharmacokinetic drug interaction potential of kratom with drugs of abuse*. Washington State University, 2023.

[93] Paul Broadhurst and Anthony W Nathan. Cardiac arrest in a young woman with the long qt syndrome and concomitant astemizole ingestion. *Heart*, 70(5):469–470, 1993.

[94] Diana Schmassmann-Suhijar, Roy Bullingham, Rodolfo Gasser, Jörg Schmutz, and Walter E Haefeli. Rhabdomyolysis due to interaction of simvastatin with mibefradil. *The Lancet*, 351(9120):1929–1930, 1998.

[95] Qijin Yin, Rui Fan, Xusheng Cao, Qiao Liu, Rui Jiang, and Wanwen Zeng. Deepdrug: a general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction. *Quantitative Biology*, 11(3):260–274, 2023.

[96] Santiago Vilar, Rave Harpaz, Eugenio Uriarte, Lourdes Santana, Raul Rabadan, and Carol Friedman. Drug—drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association*, 19(6):1066–1074, 2012.

[97] Jialiang Huang, Chaoqun Niu, Christopher D Green, Lun Yang, Hongkang Mei, and Jing-Dong J Han. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS computational biology*, 9(3):e1002998, 2013.

[98] Aurel Cami, Shannon Manzi, Alana Arnold, and Ben Y Reis. Pharmacointeraction network models predict unknown drug-drug interactions. *PloS one*, 8(4):e61468, 2013.

[99] Ke Han, Peigang Cao, Yu Wang, Fang Xie, Jiaqi Ma, Mengyao Yu, Jianchun Wang, Yaoqun Xu, Yu Zhang, and Jie Wan. A review of approaches for predicting drug–drug interactions based on machine learning. *Frontiers in pharmacology*, 12:814858, 2022.

[100] Callum R Lawson, Jenny A Hodgson, Robert J Wilson, and Shane A Richards. Prevalence, thresholds and the performance of presence–absence models. *Methods in Ecology and Evolution*, 5(1):54–64, 2014.

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

[102] Shichao Liu, Ziyang Huang, Yang Qiu, Yi-Ping Phoebe Chen, and Wen Zhang. Structural network embedding using multi-modal deep auto-encoders for predicting drug-drug interactions. In *2019 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pages 445–450. IEEE, 2019.

[103] Ha Young Jang, Jihyeon Song, Jae Hyun Kim, Howard Lee, In-Wha Kim, Bongki Moon, and Jung Mi Oh. Machine learning-based quantitative prediction of drug exposure in drug-drug interactions using drug label information. *NPJ Digital Medicine*, 5(1):88, 2022.

[104] David A Grimes and Kenneth F Schulz. An overview of clinical research: the lay of the land. *The lancet*, 359(9300):57–61, 2002.

[105] Brijesh Nair. Clinical trial designs. *Indian dermatology online journal*, 10(2):193–201, 2019.

[106] Gillian Gresham. Clinicaltrials. gov. *Principles and Practice of Clinical Trials*, pages 1–18, 2020.

[107] Ece Kavalci and Anthony Hartshorn. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Scientific reports*, 13(1):121, 2023.

[108] Hui C Kimko and Stephen B Duffull. *Simulation for designing clinical trials*. Marcel Dekker Incorporated, 2002.

[109] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[110] Eric Chalmers, Doug Hill, Vicky Zhao, and Edmond Lou. Prescriptive analytics applied to brace treatment for ais: a pilot demonstration. *Scoliosis*, 10:1–4, 2015.

[111] Felipe Feijoo, Michele Palopoli, Jen Bernstein, Sauleh Siddiqui, and Tenley E Albright. Key indicators of phase transition for clinical trials through machine learning. *Drug discovery today*, 25(2):414–421, 2020.

[112] Kevin Wu, Eric Wu, Michael DAndrea, Nandini Chitale, Melody Lim, Marek Dabrowski, Klaudia Kantor, Hanoor Rangi, Ruishan Liu, Marius Garmhausen, et al. Machine learning prediction of clinical trial operational efficiency. *The AAPS Journal*, 24(3):57, 2022.

[113] Stefan Harrer, Pratik Shah, Bhavna Antony, and Jianying Hu. Artificial intelligence for clinical trial design. *Trends in pharmacological sciences*, 40(8):577–591, 2019.

[114] Lendie Follett, Simon Geletta, and Marcia Laugerman. Quantifying risk associated with clinical trial termination: A text mining approach. *Information Processing & Management*, 56(3):516–525, 2019.

[115] Magdalyn E Elkin and Xingquan Zhu. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. *Scientific reports*, 11(1):3446, 2021.

[116] Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2):e188–e194, 2021.

[117] KS Vidhya, Ayesha Sultana, Naveen Kumar, and Harish Rangareddy. Artificial intelligence's impact on drug discovery and development from bench to bedside. *Cureus*, 15(10), 2023.

[118] Lalitkumar K Vora, Amol D Gholap, Keshava Jetha, Raghu Raj Singh Thakur, Hetvi K Solanki, and Vivek P Chavda. Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics*, 15(7):1916, 2023.

[119] Adam J Gormley. Machine learning in drug delivery. *Journal of Controlled Release*, 373:23–30, 2024.

[120] Daniela P Boso, Daniele Di Mascolo, Raffaella Santagiuliana, Paolo Decuzzi, and Bernhard A Schrefler. Drug delivery: Experiments, mathematical modelling and machine learning. *Computers in biology and medicine*, 123:103820, 2020.

[121] Sheng He, Leon G Leanse, and Yanfang Feng. Artificial intelligence and machine learning assisted drug delivery for effective treatment of infectious diseases. *Advanced drug delivery reviews*, 178:113922, 2021.

[122] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[123] Pauric Bannigan, Matteo Aldeghi, Zeqing Bao, Florian Häse, Alan Aspuru-Guzik, and Christine Allen. Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews*, 175:113806, 2021.

[124] Susmit Mhatre, Somanshi Shukla, Vivek P Chavda, Lakshmikanth Gandikota, and Vandana Patravale. Ai and ml for development of cell and gene therapy for personalized treatment. *Bioinformatics Tools for Pharmaceutical Drug Product Development*, pages 371–400, 2023.

[125] Yachao Dong, Ting Yang, Yafeng Xing, Jian Du, and Qingwei Meng. Data-driven modeling methods and techniques for pharmaceutical processes. *Processes*, 11(7):2096, 2023.

[126] Subrat Kumar Bhattamisra, Priyanka Banerjee, Pratibha Gupta, Jayashree Mayuren, Susmita Patra, and Mayuren Candasamy. Artificial intelligence in pharmaceutical and healthcare research. *Big Data and Cognitive Computing*, 7(1):10, 2023.

[127] Amit Gangwal, Azim Ansari, Iqrar Ahmad, Abul Kalam Azad, and Wan Mohd Azizi Wan Sulaiman. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Computers in Biology and Medicine*, 179:108734, 2024.

[128] Eike Petersen, Yannik Potdevin, Esfandiar Mohammadi, Stephan Zidowitz, Sabrina Breyer, Dirk Nowotka, Sandra Henn, Ludwig Pechmann, Martin Leucker, Philipp Rostalski, et al. Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE Access*, 10:58375–58418, 2022.

[129] Ying Zhang and Chen Ling. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, 4(1):25, 2018.

[130] Joachim Rudolph, Jeff Settleman, and Shiva Malek. Emerging trends in cancer drug discovery—from drugging the "undruggable" to overcoming resistance. *Cancer Discovery*, 11(4):815–821, 2021.

[131] Tamara Minko. Nanoformulation of brd4-degrading protac: Improving druggability to target the 'undruggable'myc in pancreatic cancer. *Trends in pharmacological sciences*, 41(10):684–686, 2020.

[132] John S Lazo and Elizabeth R Sharlow. Drugging undruggable molecular cancer targets. *Annual review of pharmacology and toxicology*, 56(1):23–40, 2016.

[133] Adrienne D Cox, Stephen W Fesik, Alec C Kimmelman, Ji Luo, and Channing J Der. Drugging the undruggable ras: Mission possible? *Nature reviews Drug discovery*, 13(11):828–851, 2014.

[134] Jonathan M Ostrem, Ulf Peters, Martin L Sos, James A Wells, and Kevan M Shokat. K-ras (g12c) inhibitors allosterically control gtp affinity and effector interactions. *Nature*, 503(7477):548–551, 2013.

[135] Norbert Berndt, Andrew D Hamilton, and Saïd M Sebti. Targeting protein prenylation for cancer therapy. *Nature Reviews Cancer*, 11(11):775–791, 2011.

[136] Keriann M Backus, Bruno E Correia, Kenneth M Lum, Stefano Forli, Benjamin D Horning, Gonzalo E González-Páez, Sandip Chatterjee, Bryan R Lanning, John R Teijaro, Arthur J Olson, et al. Proteome-wide covalent ligand discovery in native biological systems. *Nature*, 534(7608):570–574, 2016.

[137] Lydia Boike, Alexander G Cioffi, Felix C Majewski, Nathaniel J Henning, Michael D Jones, Gang Liu, Jeffrey M McKenna, John A Tallarico, Markus Schirle, Daniel K Nomura, et al. Discovery of a functional covalent ligand targeting an intrinsically disordered cysteine within myc. *Cell chemical biology*, 28(1):4–13, 2021.

[138] Quinlan L Sievers, Georg Petzold, Richard D Bunker, Aline Renneville, Mikołaj Słabicki, Brian J Liddicoat, Wassim Abdulrahman, Tarjei Mikkelsen, Benjamin L Ebert, and Nicolas H Thomä. Defining the human c2h2 zinc finger degrome targeted by thalidomide analogs through crbn. *Science*, 362(6414):eaat0572, 2018.

[139] Xinlin Du, Oleg A Volkov, Robert M Czerwinski, HuiLing Tan, Carlos Huerta, Emily R Morton, Jim P Rizzi, Paul M Wehn, Rui Xu, Deepak Nijhawan, et al. Structural basis and kinetic pathway of rbm39 recruitment to dcaf15 by a sulfonamide molecular glue e7820. *Structure*, 27(11):1625–1633, 2019.

[140] Kusal TG Samarasinghe and Craig M Crews. Targeted protein degradation: A promise for undruggable proteins. *Cell chemical biology*, 28(7):934–951, 2021.

[141] Donna N Petersen, Julie Hawkins, Wanida Ruangsiriluk, Kimberly A Stevens, Bruce A Maguire, Thomas N O'Connell, Benjamin N Rocke, Markus Boehm, Roger B Ruggeri, Tim Rolph, et al. A small-molecule anti-secretagogue of pcsk9 targets the 80s ribosome to inhibit pcsk9 protein translation. *Cell chemical biology*, 23(11):1362–1371, 2016.

[142] Katherine Deigan Warner, Christine E Hajdin, and Kevin M Weeks. Principles for targeting rna with drug-like small molecules. *Nature reviews Drug discovery*, 17(8):547–558, 2018.

[143] Anna C Groner, Myles Brown, et al. Role of steroid receptor and coregulator mutations in hormone-dependent cancers. *The Journal of clinical investigation*, 127(4):1126–1135, 2017.

[144] Jessica J Lin and Alice T Shaw. Resisting resistance: targeted therapies in lung cancer. *Trends in cancer*, 2(7):350–364, 2016.

[145] Philippe L Bedard, David M Hyman, Matthew S Davids, and Lillian L Siu. Small molecules, big impact: 20 years of targeted therapy in oncology. *The Lancet*, 395(10229):1078–1088, 2020.