International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Integrating Machine Learning For Risk Prediction and Adaptive Strategy in Drug Development Programs

George Stephen

Abstract

The future of clinical development is on the verge of a major transformation due to the convergence of significant new digital data sources, computing power to identify clinically meaningful patterns in the data using efficient artificial intelligence and machine-learning algorithms, and regulators embracing this change through new collaborations. This perspective summarizes insights, recent developments, and recommendations for infusing actionable computational evidence into clinical development and health care from the academy, the biotechnology industry, nonprofit foundations, regulators, and technology corporations. Analysis and learning from publically available biomedical and clinical trial data sets, real-world evidence from sensors, and health records by machine-learning architectures are discussed. Strategies for modernizing the clinical development process by integrating AI and ML based digital methods and secure computing technologies through recently announced regulatory pathways at the United States Food and Drug Administration are outlined. We conclude by discussing applications and the impact of digital algorithmic evidence on improving medical care for patients.

INTRODUCTION

Clinical drug development has remained relatively unchanged for the last 30 years. This is due, in part, to uncertainties in regulatory requirements, risk aversion, and skepticism about rapidly emerging, yet largely unproven, technologies (such as machine learning and wireless health monitoring devices and sensors), and the lack of relevant actionable biomedical data sources and advanced analytics to generate hypotheses that could motivate the development of innovative diagnostics and therapies. Testing new biomedical treatments for safety and efficacy will also require new strategies since it has been shown that existing therapies often only work for a small number of indicated individuals. The application of emerging digital technologies, such as next-generation sequencing, has increased our understanding of disease mechanisms in a larger pool of patients and the potential for developing personalized therapies. For example, the majority of the new molecular entities approved by the U.S. FDA in recent years were designed to target specific aberrations implicated in disease initiation and maintenance—a hallmark of precision medicine—which aims to tailor interventions based on individual characteristics of patients. In this light, an emerging strategy based on co-developing precision diagnostics and therapeutic agents as companion diagnostics may produce highly effective drugs with clinical outcomes that greatly exceed standard therapies. Another key challenge in the clinical development process is linked to reporting the results of most conventional clinical trials of average treatment effects that may not easily translate into making individualized treatment decisions at the routine point of care. Promising approaches to overcoming this challenge are

more streamlined processes, exploiting new digital clinical endpoints and treatment response biomarkers amenable to close and efficient monitoring (such as circulating tumor DNA), improving safety and efficacy while reducing toxicity and adverse events, and greater insights into the patient journey via sensors and low-cost imaging. Securing, standardizing, and enhancing routinely collected EHR data as a source of credible medical evidence based on RWD can facilitate the organization of clinical trials at the point of care and should improve the clinical development process. Machine learning and computer vision have enhanced many aspects of human visual perception to identify clinically meaningful patterns, e.g., imaging data and neural networks are used for various tasks ranging from medical image segmentation to generation, classification, and prediction of clinical data sets. Broadly, academic research labs, biotechnology corporations, and technology companies have been exploring the use of AI and ML in three key areas:

1. machine-based learning to predict pharmaceutical properties of molecular compounds and targets for drug discovery
2. using pattern recognition and segmentation techniques on medical images (from, e.g., retinal scans, pathology slides, and body surfaces, bones, and internal organs) to enable faster diagnoses and tracking of disease progression and generative algorithms for computational augmentation of existing clinical and imaging data sets
3. Developing deep-learning techniques on multimodal data sources, such as combining genomic and clinical data to detect new predictive models.

Despite these propositions for using ML to accelerate medical research, very few successful use cases have emerged. These limited successes have been attributed to, among other things, insufficient time elapsing since the introduction of relevant technologies and deficiency of current computer science deep learning and related ML models to generalize more complex and realistic medical data sets and tasks. Other important factors that impede the adoption of AI/ML techniques in therapeutic development include the paucity of high-quality labeled data, nascent regulations, and ethical and legal concerns about data sharing. Alternative learning systems that leverage the human brain and its neocortex and learn from fewer examples have been proposed as alternatives to deep learning but have not been widely adopted. Recently, perspectives and commentaries highlighting applications of DNN to imaging data sets, pharmaceutical properties of compounds, clinical diagnoses and genomics, computer vision applications for medical imaging, and applications of Natural Language Processing to EHR have been published. These predominantly focused on data in primary care or hospital ecosystems and early drug discovery applications and did not describe use cases and regulatory frameworks derived from a multi-stakeholder perspective for the successful embedding of AI and ML and RWE into the process of clinical development outlined in this perspective. From March 2017 to December 2018, a series of six broad, cross-institutional workshops were convened at The MIT Media Lab to discuss the current state of AI and ML and RWE usage in clinical development opportunities, challenges, and ways of addressing challenges. Participation was designed to be a multidisciplinary and multi-stakeholder, involving leading researchers from academic institutions, leaders from biopharma firms, foundations, technology corporations, and regulators to engender a broad outlook and cross-functional perspectives. Each two-part workshop was structured as follows: a series of talks outlining current challenges and opportunities and regulatory insights for introducing AI and ML in the clinical development process either as researchers or adopters, followed by

a brainstorming session with breakaway groups focusing on specific themes. This manuscript, a consolidated viewpoint on the infusion of AI and ML in clinical development, is one of the key outputs of the workshop. We focus on three key themes discussed in the workshops related to the development of next-generation medicines by the adoption of digital evidence generated by AI and ML:

(1) validation and modernizing the clinical trials process,

(2) strategies for rational use of AI and ML driven learning from real-world data and evidence and,

(3) Required regulatory oversight for integration, explanation, and derisking of AI/ML digital analytics in medical care to patients. A glossary is provided as Supplementary Material for an explanation of key terms.

2. APPLICATIONS OF ML IN DRUG DISCOVERY

The process of discovering effective new drugs is time-consuming and predominantly the most challenging part of drug development. With the advantages of learning from data, discerning patterns, and making intelligent decisions, ML-based approaches have emerged as versatile tools that can be applied in multiple stages of drug discovery, including drug design, drug screening, drug repurposing, and chemical synthesis (**Figure 1**). Moreover, considerable efforts are dedicated to developing models, tools, software, and databases based on the core architecture of ML algorithms to counter the inefficiencies and uncertainties inherent in traditional drug development methods.

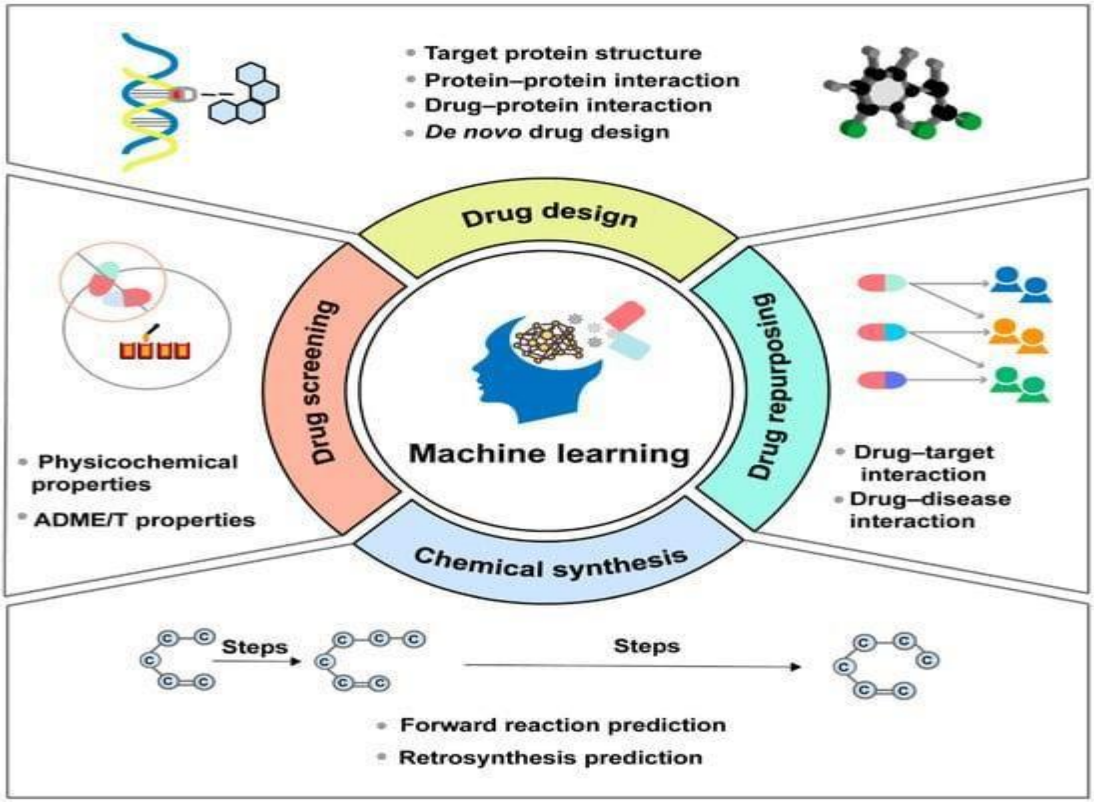

Figure 1. Machine learning can be applied in multiple stages of the drug discovery process, including drug design, drug screening, drug repurposing, and chemical synthesis

2.1. Application of ML in Drug Design

2.1.1. Prediction of the Target Protein Structure

Since proteins play crucial roles in various biological processes, their dysfunctions can lead to abnormal cell behavior and lead to the development of diseases. For selective targeting of diseases, small-molecule compounds are generally designed based on the three-dimensional (3D) chemical environment surrounding the ligand-binding sites of the target protein. Hence, predicting the 3D structure of the target protein is of great significance for structure-based drug discovery. Homology modeling has traditionally been used for this purpose, relying on known protein structures as templates. Comparatively, ML-based approaches have shown great promise in predicting the 3D structures of target proteins with improved accuracy and efficiency. For example, AlphaFold is a state-of-the-art protein structure prediction system developed by DeepMind, a leading AI company. Based on deep neural network (DNN), it has achieved remarkable success in multiple protein structure prediction competitions, demonstrating its ability to accurately predict the 3D structures of proteins by analyzing the adjacent amino acid distances and peptide bond angles. Notably, AlphaFold has significantly advanced the field of protein structure prediction and has the potential to revolutionize drug discovery. Therefore, ML-based approaches hold great potential to enhance our understanding of protein structures. It should be noted that protein structures can undergo changes in different environments, and proteins may form multiple coexisting structures under the same conditions. This complexity adds to the challenges of structure prediction.

2.1.2. Prediction of PPIs

In most cases, proteins rarely implement their functions alone but rather cooperate with other proteins to form intricate relationships known as the protein-protein interaction (PPI) network. PPIs possess indispensable functions in diverse biological processes. They can contribute to altering protein specificity, regulating protein activity, and generating novel binding sites for effector molecules. Hence, understanding and targeting PPIs offers opportunities to design innovative drugs to modulate complex biological processes.

Currently, ML-based methods for PPI prediction can be broadly grouped into structure-based and sequence-based categories. Structure-based approaches mainly leverage the knowledge of protein structure similarity to predict PPIs. For example, IntPred, a random forest ML tool, was developed to predict protein-protein interface sites based on structural features. Compared with other methods, the IntPred predictor showed strong performance in identifying interactions at both the surface-patch and residue levels on independent test sets of both obligate and transient complexes (Matthews' Correlation Coefficient (MCC) = 0.370, accuracy = 0.811, specificity = 0.916, sensitivity = 0.411). Struct2Graph, a graph attention network (GAT)-based classifier, was proposed to identify PPIs directly from the 3D structure of protein chains. The accuracy of Struct2Graph on balanced sets with equal numbers of positive and negative pairs was 0.9989, and the average accuracy of five-fold cross-validation on unbalanced sets with a ratio of positive and negative pairs of 1:10 was 0.9942. Comparatively, sequence-based PPI prediction approaches aim to identify physical interactions between two proteins by leveraging information from their protein sequences. DNNs provide a robust solution for this purpose. They comprise multiple layers of interconnected neurons, allowing them to extract complex patterns and features from data automatically. For example, DeepPPI applied DNNs to effectively learn protein representations from common protein descriptors, thereby contributing to PPI prediction. It can achieve excellent performance on the *S. cerevisiae* dataset with an accuracy of 0.925, precision of 0.9438, recall of 0.9056, specificity of

0.9449, MCC of 0.8508, and area under the curve (AUC) of 0.9743, respectively. Extensive experiments showed that DeepPPI could learn the useful features of protein pairs through a layer-wise abstraction, resulting in better predictive performance than existing methods on core *S. cerevisiae*, *H. pylori*, and *H. sapiens* datasets. In addition, based on the Uniprot database, Li et al. developed a DELPHI, a new sequence-based deep ensemble model for PPI-binding sites' prediction. Therefore, ML-based approaches have great potential in enhancing the identification of PPI sites. Compared with sequence-based approaches, structure-based ones are limited by the scarcity of available protein structures and the low quality of familiar protein structures.

2.1.3. Prediction of DTIs

Most drugs exert therapeutic effects by specifically interacting with target molecules within the body, such as enzymes, receptors, and ion channels. Hence, the accurate prediction of DTIs is a pivotal step in the drug design pipeline. As traditional experimental approaches are time-consuming and costly, researchers have increasingly developed and applied ML-based methods to predict DTIs. These methods primarily focus on three key aspects: predicting the binding sites of drugs on target molecules, estimating the binding affinity between drugs and targets, and determining the binding pose or conformation of the drug within the target molecule.

Firstly, binding sites, also called binding pockets, are specific locations within a protein where interactions occur between the protein and a ligand (such as a drug molecule). By introducing a deep convolutional neural network (CNN), Cui et al. developed a sequence-based method, DeepC-SeqSite, for predicting protein-ligand binding residues. Notably, this method exhibited superior performance compared with multiple existing sequence-based and 3D- 3D-structure-based methods, including the leading ligand-binding method, COACH. Similarly, Zhou et al. proposed a binding site prediction method called AGAT-PPIS based on augmented GAT. It demonstrated significant improvements over the state-of-the-art method, achieving an accuracy increase of 8% on the benchmark test set. Moreover, binding affinity represents the strength of an interaction between a drug and its target. Some tools based on ML and DL algorithms have been applied to determine DTIs' binding affinity, such as DEELIG and GraphDelta. In addition, the active conformation of ligands plays a crucial role in facilitating the effective binding between proteins and drugs. By combining random forest and CNN strategies, Nguyen et al. proposed a scoring function to select the most relevant poses generated by docking software tools, including GOLD, GLIDE, and Autodock Vina, thereby contributing to obtaining more accurate and effective ligand-target binding configurations. Therefore, ML algorithms have been extensively employed to predict DTIs and hold the potential to facilitate the design of new drugs.

2.1.4. De Novo *Drug Design*

De novo drug design refers to creating new drug molecules from scratch using computational methods without relying on existing bioactive compounds or known drug structures. It involves designing molecules with specific properties and functions targeting a particular disease or condition. Compounds developed with traditional *de novo* drug design methods (e.g., the fragment-based approach) usually have poor drug metabolism and pharmacokinetics properties. They are hard to synthesize due to the complexity and impracticality of compound structures. Therefore, there is a high demand for new methods to explore chemical entities that meet the requirements of biological activity, drug metabolism, pharmacokinetics, and synthesis practicality.

Recently, ML-based approaches, especially auto-encoder variants (e.g., the variational auto-encoder (VAE) and adversarial auto-encoder (AAE)), have gained attention in the field of *de novo* drug design. PaccMann^{RL} is an example of these approaches combining a hybrid VAE with reinforcement learning for the *de novo* design of anti-cancer molecule design from transcriptomic data. Similarly, another approach, druGAN, utilizes a deep generative AAE model to generate novel molecules with specific anti-cancer properties. In addition, a Wasserstein GAN and GCN-based model, known as MedGAN, has been successfully developed to generate novel quinoline-scaffold molecules from complicated molecular graphs and evaluate drug-related properties. It has been demonstrated that the MedGAN produced 25% effective molecules, 62% fully connected, 92% are quinoline, 93% are novel, and 95% are unique. Coley et al. defined a synthetic complexity score, SCScore, that utilizes precedent reaction knowledge to train a neural network model for evaluating the level of synthetic complexity to address the difficulty in synthesizing generated molecules. Therefore, ML- ML-empowering approaches play crucial roles in *de novo* drug design, revolutionizing the discovery and development of new drugs.

2.2. Application of ML in Drug Screening

2.2.1. Prediction of the Physicochemical Properties

The physicochemical properties of drugs, mainly including solubility, ionization degree, partition coefficient, permeability coefficient, and stability, play a significant role in determining their behavior (e.g., bioavailability, absorption, transportation, and permeability) in biological systems as well as the environment, and in evaluating their potential risks to human health. Hence, these properties are assessed during drug screening to select promising candidates for further development and optimization. Multiple ML-based tools have been proposed to predict the physicochemical properties of molecules. For example, Francoeur et al. developed a molecule attention Transformer called SolTranNet to predict aqueous solubility from the SMILES representation of drug molecules. It has been demonstrated to function as a classifier for filtering insoluble compounds, achieving a sensitivity of 0.948 on Challenge to Predict Aqueous Solubility (SC2) datasets, which is competitive with other methods. Moreover, by using molecular fingerprints and four ML algorithms, Zang et al. developed a quantitative structure-property relationship workflow to predict six physicochemical properties of environmental chemicals, such as water solubility, octanol-water partition coefficient, melting point, boiling point, bioconcentration factor, and vapor pressure [59]. Therefore, these ML-based predictors are valuable tools in drug discovery, as they can help screen potential drug candidates based on their physicochemical properties.

2.2.2. Prediction of the ADME/T Properties

Once hit or lead compounds are identified during the drug discovery, tests, and evaluations are conducted to assess their absorption, distribution, metabolism, excretion, and toxicity (ADME/T) properties. These pharmacokinetic properties are essential for understanding how the compounds will behave in the human body and whether they have the potential to be as safe and effective as drugs. Imbalanced ADME/T properties frequently cause the failure of drug candidates in the late stages of drug development and may even lead to the withdrawal of approved drugs. Hence, ADME/T properties are often employed as molecular filters to screen large databases of compounds in the early stage of drug discovery, thereby helping to increase efficiency and improve the success rate of drug screening.

To detect the ADME/T properties of drugs, various evaluation criteria such as hepatotoxicity, passing through the blood-brain barrier (BBB), plasma protein binding (PPB), and cytochrome P450 2D6

(CYP2D6) inhibition are commonly used. Accordingly, growing interest has been in developing ML-based tools for predicting these criteria. For example, Tian et al. developed a web server called ADMETboost that utilizes the powerful extreme gradient boosting (XGBoost) model to learn about molecule features from multiple fingerprints and descriptors, allowing for the accurate prediction of ADME/T properties, such as Caco2, BBB, CYP2C9 inhibition, CL-Hepa, and hERG. It has been demonstrated that this model can achieve remarkable results in the Therapeutics Data Commons ADMET benchmark, ranking first in 18 out of 22 tasks and within the top three in 21 tasks. Similarly, by utilizing more than 13 000 compounds obtained from the PubChem BioAssay Database, Li et al. proposed a multitask autoencoder DNN model to predict the inhibitors of five major cytochrome P450 (CYP450) isoforms (1A2, 2C9, 2C19, 2D6 and 3A4). Especially the multi-task DNN model achieved average prediction accuracies of 86.4% in 10-fold cross-validation and 88.7% on external test datasets, outperforming single-task models, earlier described classifiers, and conventional ML methods.

Furthermore, the Tox21 Challenge is a collaborative effort aimed at developing predictive models for toxicity assessment using high-throughput screening data. In this context, Mayr et al. developed a DL pipeline, DeepTox, for toxicity prediction. It outperformed all other computational methods (e.g., naïve Bayes, random forest, and support vector machine) in 10 out of 15 cases in the Tox21 Challenge. Therefore, ML algorithms have made significant progress in predicting the ADME/T properties of drugs, contributing to guiding drug safety assessment and preclinical research.

2.3. Application of ML in Drug Repurposing

Drug repurposing, or drug repositioning, is a strategy to identify new indications from approved or investigational (including failed clinical trials) that have not been approved. As this approach takes advantage of the extensive safety testing conducted during clinical trials for other purposes, repurposing known drugs speeds up drug development. It presents cost-saving advantages compared to developing entirely new drugs from scratch. Currently, researchers are increasingly developing and applying ML-based methods to conduct drug repurposing, which can be broadly divided into target-centered and disease-centered approaches.

In target-centered drug repurposing, network-based methods have been widely applied to search for new targets for known drugs. For example, by employing autoencoder and Positive-unlabeled matrix completion algorithms, Zeng et al. developed a calculation method called deepDTnet to identify new targets for known drugs from a heterogeneous drug–gene-disease network. Experiments have shown that the deepDTnet achieved a high accuracy in predicting new targets of existing drugs (AUC = 0.963), which is superior to traditional ML methods. Similarly, by combining the network diffusion algorithm and the dimensionality reduction approach, Luo et al. developed DTINet, a novel network-integration procedure for DTI prediction and drug repositioning. It can outperform existing methods, with AUC and area under precision-recall (AUPR) 5.7% and 5.9% higher than the second-best method, respectively, providing an effective drug discovery and target identification tool.

In addition, disease-centered approaches mainly aim to identify drug-disease relationships and can be widely divided into similarity-based and network-based ones. Similarity-based methods have achieved significant progress by combining drug or disease characteristics with the known drug-disease associations. For example, based on the assumption that similar drugs are commonly associated with similar diseases, Luo et al. proposed a novel computational approach called MBiRW, which combines

similarity measurements and a Bi-Random walk algorithm to recognize potential novel indications for a specific drug. MBiRW can achieve high accuracy in predicting drug-disease associations (AUC = 0.917), which is superior to other methods. In addition, network-based methods integrate information from different biological networks to improve the predictive accuracy of drug-disease relationships. For example, Doshi et al. developed a graph neural network model called GDRnet for drug repurposing, which can efficiently screen existing drugs in the database and predict their unknown therapeutic effects by evaluating the scores of drug-disease pairs. Therefore, ML technology holds significant promise in drug repurposing, providing strong support for accelerating drug discovery.

2.4. Application of ML in Chemical Synthesis

Organic synthesis is a key part of the small-molecule drug discovery process. New molecules are synthesized along the path of compound optimization to achieve improved properties. To promote molecule synthesis, researchers have developed multiple ML-based computational tools applicable to retrosynthesis prediction and forward reaction prediction.

2.4.1. Retrosynthesis Prediction

Retrosynthesis planning aims to identify efficient synthetic routes for a desired molecule by recursively converting it into easier precursors. Therefore, it can effectively solve the synthesis of complex molecules to facilitate the development of organic synthesis science. Several ML-based approaches, including template-based and template-free approaches, have been used for retrosynthesis planning.

The template-based approach systematically compares the target molecule with a set of templates, each representing alternative substructure patterns during a chemical reaction. Segler et al., published in Nature, presented the first work involving DNNs for this issue. They found that Monte Carlo tree search (MCTS) combined with DNNs and symbolic rules can be utilized to perform chemical synthesis effectively. The routes generated by the model were comparable to those reported in the literature in a double-masked AB test, thereby confirming the model's accuracy. However, it is worth noting that template-based approaches cannot be extended beyond templates, limiting their predictive ability.

The template-free method aims to uncover hidden relationships within the data concerning reaction mechanisms rather than relying on direct matching. For example, using neural sequence-to-sequence models, Liu et al. proposed the template-free method, called seq2seq, to perform the retrosynthetic reaction-prediction tasks. This model was based on an encoder-decoder framework consisting of two recurrent neural networks (RNNs) and was trained on a dataset of 50,000 experimental reactions extracted from the United States patent literature, demonstrating comparable performances to the rule-based expert system model. Therefore, ML algorithms have been extensively employed to conduct retrosynthetic analysis and hold the potential to facilitate chemical synthesis.

2.4.2. Forward Reaction Prediction

Contrary to retrosynthesis analysis, forward reaction prediction aims to identify potential molecules that can be synthesized from given reactants and reagents. Given the reactant molecules as input, the ML model analyzes their structural and chemical properties to generate predictions about the resulting products and reaction conditions. For example, Wei et al. introduced a novel reaction fingerprinting approach that utilizes neural networks to predict reaction types. The prediction results of this method on 16 essential reactions of alkyl halides and alkenes indicate that neural networks can contribute to identifying key

features from the structure of reactant molecules to classify new reaction types. Similarly, Coley et al. proposed a neural network model to predict the main products of chemical reactions by training the data extracted from a collection of 150,000 compounds' reaction templates in the US patent database. In addition, reaction conditions (e.g., solvent and temperature) are critical in practical chemical synthesis reactions to maximize desired product yield. Based on this, Gao et al. proposed a neural network model to predict the optimal reaction conditions for various reactions. This model was trained using a vast dataset of nearly 10 million entries extracted from the Reaxys database and can effectively predict the ideal catalyst, solvent, reagent, and temperature for a given reaction, facilitating the optimization of reaction conditions. Therefore, the utilization of ML-based models can assist in predicting reaction types, accelerating the discovery of new chemical molecules, and identifying optimal reaction conditions, thereby holding great potential in improving the efficiency of chemical synthesis processes.

3. OPPORTUNITIES FOR TRANSFORMER-BASED ML MODELS IN EMPOWERING DRUG DISCOVERY

The Transformer model, first proposed in the paper ‘Attention is All You Need’ by Vaswani et al., is a highly advanced DL architecture utilizing self-attention mechanisms. As it allows for parallelization and captures long-range dependencies more efficiently than traditional RNN models, the Transformer model has proven highly effective in many tasks and has set new benchmarks in the corresponding fields. Given its advantages, it has emerged as a promising future direction of ML in drug discovery (**Figure** 2).

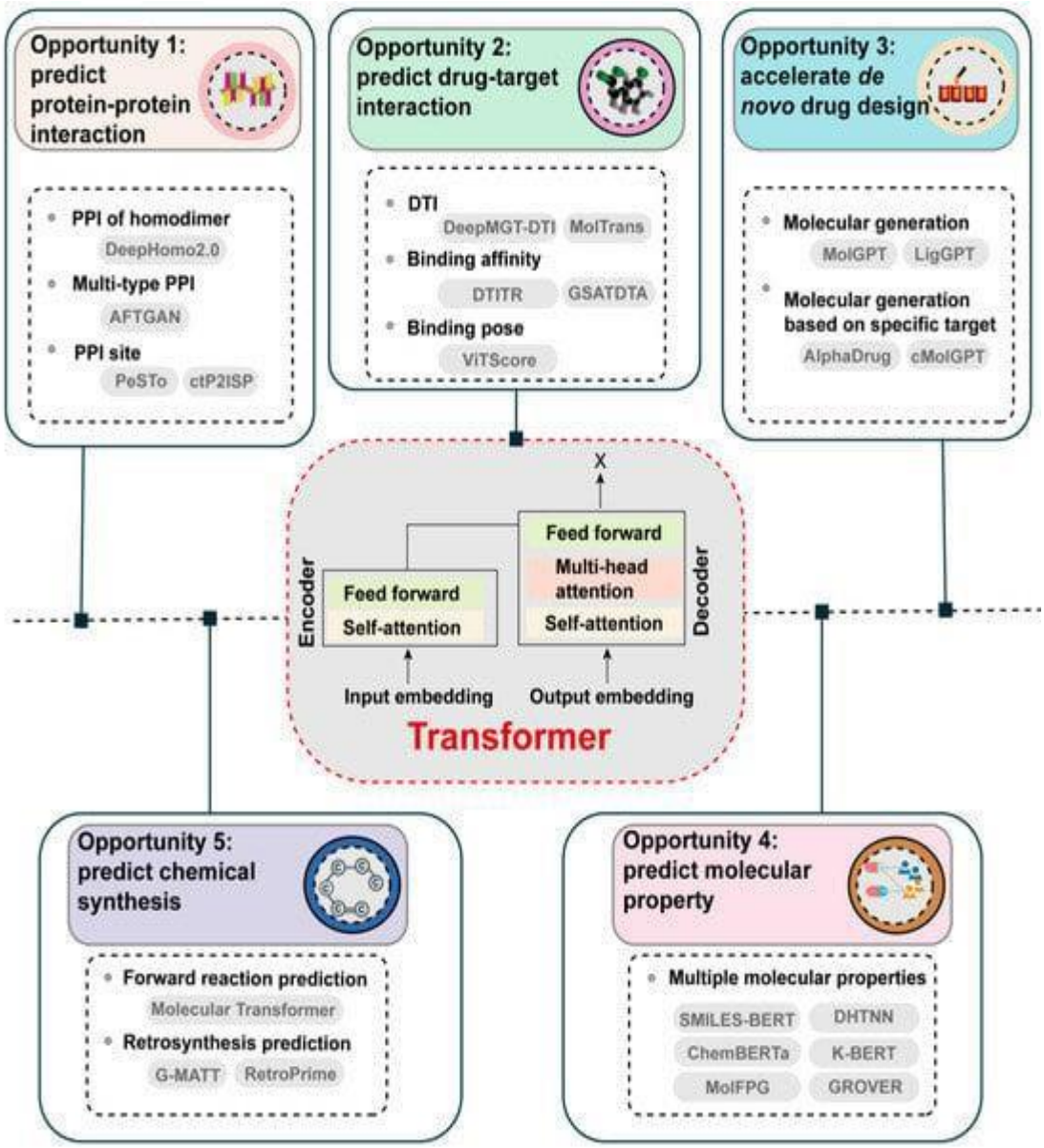

Figure 3. Opportunities for Transformer-based models in empowering drug discovery

3.1. Opportunity 1: Transformer Models Empowering PPIs Identification

Existing ML-based approaches mainly use CNNs to extract low-dimensional features from protein sequences based on the amino acid composition while disregarding the long-range relationships within these sequences. Fortunately, transformers can capture the long-distance dependencies in the protein sequences, making them suitable for predicting whether and how given proteins interact. For example, by utilizing the advantage of the Transformer model in evolutionary scale modeling-multiple sequence alignment, Lin et al. developed DeepHomo2.0, a DL-based model that predicts PPIs of homodimeric complexes by combining Transformer features, monomer structure information, and direct-coupling analysis. The results showed that DeepHomo2.0 can achieve a high accuracy of over 70% and 60% in terms of experimental monomer structure and predicted monomer structure for the top 10 contacts predicted on the Protein Data Bank (PDB)test set, respectively, which is superior to the DCA-based, protein language model-based and other ML-based methods. Similarly, Kang et al.proposed AFTGAN, a

neural network that combines Transformer and GAT frameworks for effective protein information extraction and multi-type PPI prediction. Experimental comparisons validated the superior performance of AFTGAN in accurately predicting the PPIs of unknown proteins. Therefore, given the advantage of the Transformer in extracting protein sequences, it has demonstrated remarkable potential in advancing the prediction of PPIs.

3.2. Opportunity 2: Transformer Models Empowering DTIs' Identification

Despite the remarkable performance improvement of DL models in DTI prediction, the primary challenge lies in the limited representation of drugs in these methods, as they only consider SMILES sequences, SMARTS strings, or molecular graphs, failing to capture comprehensive drug representations. It is worth noting that Transformers can be employed either independently or in combination with other AI algorithms to address these problems. For example, DeepMGT-DTI, a DL model incorporating a Transformer network and multilayer graph information, can effectively capture the structural features of drugs, leading to improved DTI prediction. Experiments have demonstrated that the DeepMGT-DTI can achieve an AUC of 90.24%, an AUPR of 77.11%, an F1 score of 79.31%, and an accuracy of 85.15% on the DrugBank dataset. These performance metrics surpassed those of previous target sequence-structure models, such as Deep DTA and TransformerCPI. Moreover, GSATDTA, a novel triple-channel model based on graph–sequence attention and Transformer, has been developed to predict the drug-target binding affinity with outstanding performance. Therefore, transformer models have shown promising results for the prediction of DTIs.

3.3. Opportunity 3: Transformer Models Empowering De Novo Drug Design

Most existing deep generative models either focus on virtual screening of the available database of compounds by DTI binding-affinity prediction or unconditionally generate molecules with specific physicochemical and pharmacological properties, ignoring protein targets' function during the generation process. In contrast, Transformer models can consider the protein target and achieve target-specific molecular generation. For example, AlphaDrug, a method for protein target-specific *de novo* drug design, has been recently proposed. It utilizes a modified Transformer to optimize the learning of protein information and integrates an efficient MCTS guided by the Transformer's predictions and docking values. Notably, in terms of average docking score, uniqueness, the octanol-water partition coefficient logP, the quantitative estimate of drug-likeness (QED), synthetic accessibility (SA), and Natural products- likeness (NP-likeness) criteria, AlphaDrug is superior to other methods (such as LiGANN, SBMolGen, and SBDD-3D). In addition, the GPT model is a powerful language generation model that can be fine-tuned for specific tasks after pre-training on large amounts of text data. It has been successfully applied to accelerate molecular generation for specific targets in the field of drug discovery. For example, cMolGPT, a GPT-inspired model, is useful for target-specific *de novo* molecular generation. The chemical space of the compounds generated by cMolGPT closely matches that of real target-specific ones.

3.4. Opportunity 4: Transformer Models Empowering Molecular Property Prediction

Despite the widespread application of ML-based models, the shortage of labeled data remains a significant challenge in inefficient molecular property predictions. To address this, researchers are exploring unlabeled data and leveraging transformer-based self-supervised learning (e.g., BERT) to improve predictions on small-scale labeled data. Currently, several BERT-based pre-training methods for molecular property prediction have been proposed. For example, a novel pre-training method, K-BERT,

was developed to extract chemical information from SMILES similar to chemists for molecular property prediction in drug discovery. The K-BERT model exhibited superior performance in 8 out of 15 tasks, thus reflecting the efficacy and benefits of the proposed pre-training approach in drug discovery. Specifically, K- BERT had an average AUC score of 0.806, outperforming other competing methods (e.g., XGBoost-MACCS, XGBoost-ECFP4, HRGCN+, and Attentive FP). Moreover, Wang et al. proposed a two-stage (pre-training and fine-tuning) model called SMILES-BERT that could use both unlabeled data and labeled data to improve molecular property prediction. Compared with a range of state-of-the-art approaches (e.g., CircularFP, NeuralFP, Seq2seqFP, Seq3seqFP), it exhibited superior performance on three different datasets (the LogP dataset, PM2 dataset, and PCBA-686978 dataset) with accuracies of 0.9154, 0.7589, and 0.8784, respectively. Therefore, these Transformer-based predictors are essential tools for molecular property prediction, contributing to efficiently screening potential drug candidates.

3.5. Opportunity 5: Transformer Models Empowering Chemical Synthesis

Previous sequence-based approaches commonly employed RNNs for the encoder and decoder, with a single-head attention layer connecting them. These models treated reactants and reagents separately in the input by utilizing atom mapping, which limits the interpretability of the model. Fortunately, Transformer-powered models have shown the potential to accelerate chemical synthesis. One notable example is the effectiveness of the multi-head attention Molecular Transformer model in predicting chemical reactions and reaction conditions. In addition, inspired by the success of the Molecular Transformer for forward reaction prediction, Schwaller et al. proposed an enhanced Molecular Transformer architecture coupled with a hyper-graph exploration algorithm for automated retrosynthetic pathway prediction. This approach surpasses previous ML-based methods by not only predicting reactants but also identifying reagents for each retrosynthetic step, thereby significantly raising the complexity of the prediction task.

4.0 DRUG DISCOVERY THROUGH AI/ML

Many pharmaceutical corporations have invested resources in this area because of the possibility of integrating machine-learning models through all the phases of drug discovery. The chances of this report disallow a detailed analysis of this action. ML is being used on these datasets in genomics for a variety of reasons, including defining disease subtypes, finding disease biomarkers, drug discovery and repurposing, and medication response prediction.

Many large pharmaceutical businesses work on AI-related research and development programs or collaborations. AstraZeneca and Benevolent, for example, are using AI to speed up the discovery of new potential drug targets by combining genomes, chemistry, and clinical data. GlaxoSmithKline (GSK) has invested in the biotechnology company 23andMe, acquiring entry to The company's datasets use machine learning to discover pharmacological targets. The drugmaker has also developed collaborations with AI drug discovery businesses. An additional area of therapeutic research aided by machine learning is genome editing, which involves removing, adding, or altering parts of DNA. The advent of targeted treatment has made growth in precision medicine. Genome-editing techniques are increasingly employed for therapeutic purposes, such as replacing or altering a faulty gene in patients. The study better understands the significance of genes and DNA sequences. CRISPR is the most flexible, cost-effective, and straightforward technology for genome editing currently available. It is trained with ML and DL algorithms to improve its efficiency and accuracy (Fig.3).

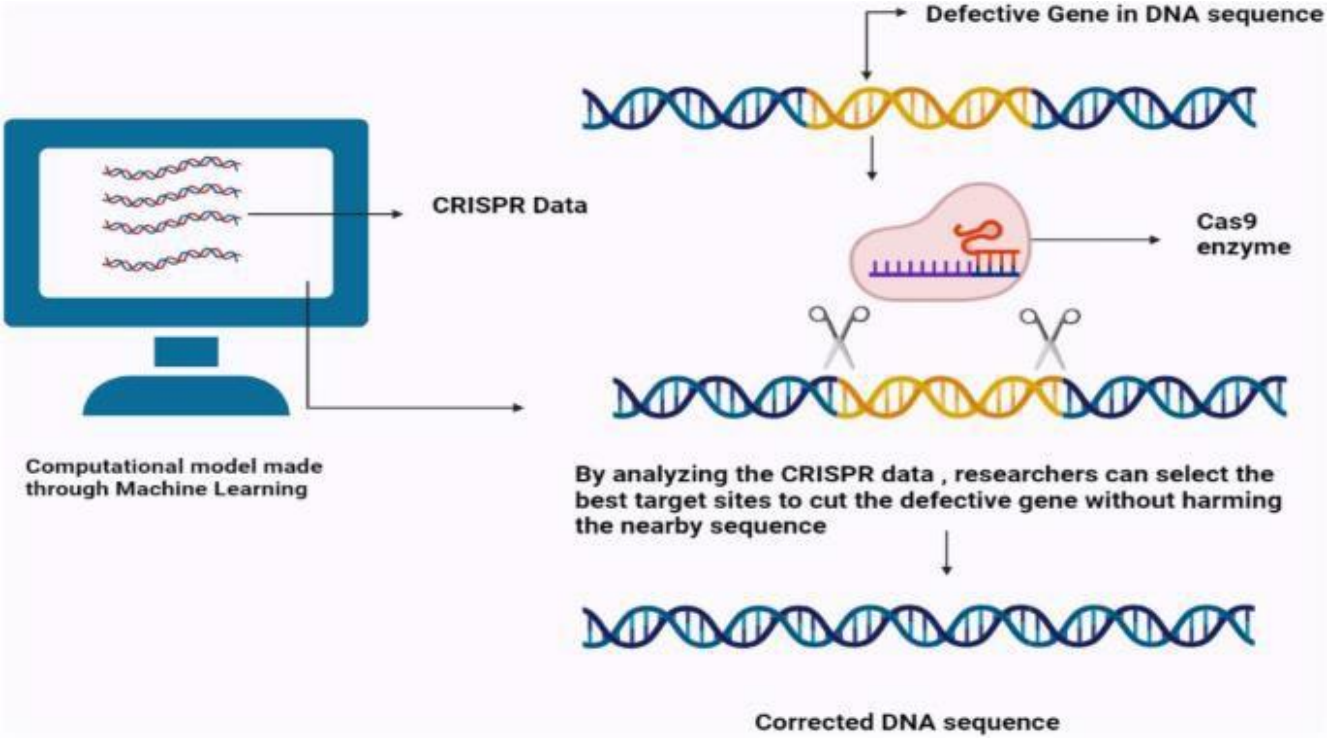

Fig. 3 A hypothetical illustration of CRISPR gene editing through a machine-learning computational model

ML algorithmic approaches have been devised to forecast the activity of the editing system, the precise differences caused by edits, and off-target consequences such as unintentional DNA alternation that might hamper the technology. Advancement in silico prediction will be critical for developing experimental disease models and speeding up and notifying the development of safer and more precise medicines.

For these reasons, pharmaceutical corporations are prioritizing CRISPR technologies. GSK has announced a multi-million-dollar agreement with the University of California to build a CRISPR laboratory, with GSK's artificial intelligence section supporting data analysis.

CONCLUSION

The research and development of new drugs can contribute to meeting the human demand for treating diseases and provide more effective, safer, and more convenient treatment options. Compared with the traditional strategies of drug discovery, ML-based approaches have the potential to reduce time and costs, improve safety, and bridge the gap between drug discovery and drug effectiveness, making them increasingly favored by the pharmaceutical industry and academia. In particular, the introduction of chatGPT has sparked researchers' growing interest and exploration in leveraging the Transformer model's NLP capabilities, particularly its self-attention mechanisms, to accelerate multiple stages of the drug discovery process, thereby opening up new opportunities for advancements.

However, the current challenges in ML-based models can generate false positives or false negatives, potentially leading to incorrect predictions and resource waste. Further in vitro and in vivo experiments, as well as clinical trials, are needed to fully demonstrate the practicability of ML-based drug discovery and

obtain more reliable and accurate results. Therefore, future research should focus on improving data quality, enhancing the interpretability of ML algorithms, and integrating them with human professional knowledge to increase the efficacy of drug discovery.

REFERENCE:

1. Monteiro, N.R.C.; Pereira, T.O.; Machado, A.C.D.; Oliveira, J.L.; Abbasi, M.; Arrais, J.P. FSM-DDTR: End-to-end feedback strategy for multi-objective De Novo drug design using transformers. *Comput. Biol. Med.* **2023**, *164*, 107285. [**Google Scholar**] [**CrossRef**] [**PubMed**]
2. Song, T.; Ren, Y.; Wang, S.; Han, P.; Wang, L.; Li, X.; Rodriguez-Patón, A. DNMG: Deep molecular generative model by fusion of 3D information for de novo drug design. *Methods* **2023**, *211*, 10–22. [**Google Scholar**] [**CrossRef**]
3. Macedo, B.; Ribeiro Vaz, I.; Taveira Gomes, T. MedGAN: Optimized generative adversarial network with graph convolutional networks for novel molecule design. *Sci. Rep.* **2024**, *14*, 1212. [**Google Scholar**] [**CrossRef**]
4. Panapitiya, G.; Girard, M.; Hollas, A.; Sepulveda, J.; Murugesan, V.; Wang, W.; Saldanha, E. Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction. *ACS Omega* **2022**, *7*, 15695–15710. [**Google Scholar**] [**CrossRef**]
5. Francoeur, P.G.; Koes, D.R. SolTranNet-A Machine Learning Tool for Fast Aqueous Solubility Prediction. *J. Chem. Inf. Model.* **2021**, *61*, 2530–2536. [**Google Scholar**] [**CrossRef**]
6. Zang, Q.; Mansouri, K.; Williams, A.J.; Judson, R.S.; Allen, D.G.; Casey, W.M.; Kleinstreuer, N.C. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *J. Chem. Inf. Model.* **2017**, *57*, 36–49. [**Google Scholar**] [**CrossRef**]
7. Tian, H.; Ketkar, R.; Tao, P. ADMETboost: A web server for accurate ADMET prediction. *J. Mol. Model.* **2022**, *28*, 408. [**Google Scholar**] [**CrossRef**]
8. Schyman, P.; Liu, R.; Desai, V.; Wallqvist, A. vNN Web Server for ADMET Predictions. *Front. Pharmacol.* **2017**, *8*, 889. [**Google Scholar**] [**CrossRef**] [**PubMed**]
9. Wei, Y.; Li, S.; Li, Z.; Wan, Z.; Lin, J. Interpretable-ADMET: A web service for ADMET prediction and optimization based on deep neural representation. *Bioinformatics* **2022**, *38*, 2863–2871. [**Google Scholar**] [**CrossRef**] [**PubMed**]
10. Deng, D.; Chen, X.; Zhang, R.; Lei, Z.; Wang, X.; Zhou, F. XGraphBoost: Extracting Graph Neural Network-Based Features for a Better Prediction of Molecular Properties. *J. Chem. Inf. Model.* **2021**, *61*, 2697–2705. [**Google Scholar**] [**CrossRef**] [**PubMed**]
11. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **2016**, *3*, 80. [**Google Scholar**] [**CrossRef**]
12. Li, X.; Xu, Y.; Lai, L.; Pei, J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharm.* **2018**, *15*, 4336–4345. [**Google Scholar**] [**CrossRef**]
13. Shaker, B.; Yu, M.S.; Song, J.S.; Ahn, S.; Ryu, J.Y.; Oh, K.S.; Na, D. LightBBB: Computational prediction model of blood-brain-barrier penetration based on LightGBM. *Bioinformatics* **2021**, *37*, 1135–1139. [**Google Scholar**] [**CrossRef**]
14. Tang, Q.; Nie, F.; Zhao, Q.; Chen, W. A merged molecular representation deep learning method for

blood-brain barrier permeability prediction. *Brief. Bioinform.* **2022**, *23*, bbac357. [**Google Scholar**] [**CrossRef**]

15. Jang, W.D.; Jang, J.; Song, J.S.; Ahn, S.; Oh, K.S. PredPS: Attention-based graph neural network for predicting stability of compounds in human plasma. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 3532–3539. [**Google Scholar**] [**CrossRef**]
16. Khaouane, A.; Khaouane, L.; Ferhat, S.; Hanini, S. Deep Learning for Drug Development: Using CNNs in MIA-QSAR to Predict Plasma Protein Binding of Drugs. *AAPS PharmSciTech* **2023**, *24*, 232. [**Google Scholar**] [**CrossRef**] [**PubMed**]
17. Zeng, X.; Zhu, S.; Lu, W.; Liu, Z.; Huang, J.; Zhou, Y.; Fang, J.; Huang, Y.; Guo, H.; Li, L.; et al. Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* **2020**, *11*, 1775–1797. [**Google Scholar**] [**CrossRef**] [**PubMed**]
18. Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; Zeng, J. NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics* **2019**, *35*, 104–111. [**Google Scholar**] [**CrossRef**]
19. Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **2017**, *8*, 573. [**Google Scholar**] [**CrossRef**]
20. Luo, H.; Wang, J.; Li, M.; Luo, J.; Peng, X.; Wu, F.X.; Pan, Y. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* **2016**, *32*, 2664–2671. [**Google Scholar**] [**CrossRef**]
21. Doshi, S.; Chepuri, S.P. A computational approach to drug repurposing using graph neural networks. *Comput. Biol. Med.* **2022**, *150*, 105992. [**Google Scholar**] [**CrossRef**] [**PubMed**]
22. Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. deepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **2019**, *35*, 5191–5198. [**Google Scholar**] [**CrossRef**]
23. Jiang, H.J.; Huang, Y.A.; You, Z.H. Predicting Drug-Disease Associations via Using Gaussian Interaction Profile and Kernel-Based Autoencoder. *BioMed Res. Int.* **2019**, *2019*, 2426958. [**Google Scholar**] [**CrossRef**]
24. Ghorbanali, Z.; Zare-Mirakabad, F.; Salehi, N.; Akbari, M.; Masoudi-Nejad, A. DrugRep-HeSiaGraph: When heterogenous siamese neural network meets knowledge graphs for drug repurposing. *BMC Bioinform.* **2023**, *24*, 374. [**Google Scholar**] [**CrossRef**] [**PubMed**]
25. Suviriyapaisal, N.; Wichadakul, D. iEdgeDTA: Integrated edge information and 1D graph convolutional neural networks for binding affinity prediction. *RSC Adv.* **2023**, *13*, 25218–25228. [**Google Scholar**] [**CrossRef**]
26. Segler, M.H.S.; Preuss, M.; Waller, M.P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610. [**Google Scholar**] [**CrossRef**]
27. Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113. [**Google Scholar**] [**CrossRef**]
28. Thakkar, A.; Chadimová, V.; Bjerrum, E.J.; Engkvist, O.; Reymond, J.L. Retrosynthetic accessibility score (RAscore)—Rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021**, *12*, 3339–3349. [**Google Scholar**] [**CrossRef**]
29. Wei, J.N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic

Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732. [**Google Scholar**] [**CrossRef**]

30. Coley, C.W.; Barzilay, R.; Jaakkola, T.S.; Green, W.H.; Jensen, K.F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443. [**Google Scholar**] [**CrossRef**]
31. Gao, H.; Struble, T.J.; Coley, C.W.; Wang, Y.; Green, W.H.; Jensen, K.F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465–1476. [**Google Scholar**] [**CrossRef**]
32. Marcou, G.; Aires de Sousa, J.; Latino, D.A.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. Expert system for predicting reaction conditions: The Michael reaction case. *J. Chem. Inf. Model.* **2015**, *55*, 239–250. [**Google Scholar**] [**CrossRef**]
33. You, Z.H.; Li, S.; Gao, X.; Luo, X.; Ji, Z. Large-scale protein-protein interactions detection by integrating big biosensing data with computational model. *BioMed Res. Int.* **2014**, *2014*, 598129. [**Google Scholar**] [**CrossRef**]
34. Chan, H.C.S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **2019**, *40*, 592–604. [**Google Scholar**] [**CrossRef**]
35. Muhammed, M.T.; Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem. Biol. Drug Des.* **2019**, *93*, 12–20. [**Google Scholar**] [**CrossRef**]
36. Zhang, Y.; Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1029–1034. [**Google Scholar**] [**CrossRef**]
37. Tang, T.; Zhang, X.; Liu, Y.; Peng, H.; Zheng, B.; Yin, Y.; Zeng, X. Machine learning on protein-protein interaction prediction: Models, challenges and trends. *Brief. Bioinform.* **2023**, *24*, bbad076. [**Google Scholar**] [**CrossRef**]
38. Soleymani, F.; Paquet, E.; Viktor, H.; Michalowski, W.; Spinello, D. Protein-protein interaction prediction with deep learning: A comprehensive review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 5316–5341. [**Google Scholar**] [**CrossRef**]
39. Li, S.; Wu, S.; Wang, L.; Li, F.; Jiang, H.; Bai, F. Recent advances in predicting protein- protein interactions with the aid of artificial intelligence algorithms. *Curr. Opin. Struct. Biol.* **2022**, *73*, 102344. [**Google Scholar**] [**CrossRef**]
40. Dowden, H.; Munro, J. Trends in clinical success rates and therapeutic focus. *Nat. Rev. Drug Discov.* **2019**, *18*, 495–496. [**Google Scholar**] [**CrossRef**] [**PubMed**]
41. Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; Wang, F. Artificial intelligence in drug discovery: Applications and techniques. *Brief. Bioinform* **2022**, *23*, bbab430. [**Google Scholar**] [**CrossRef**] [**PubMed**]
42. Mak, K.K.; Pichika, M.R. Artificial intelligence in drug development: Present status and future prospects. *Drug Discov. Today* **2019**, *24*, 773–780. [**Google Scholar**] [**CrossRef**]
43. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26*, 80–93. [**Google Scholar**] [**CrossRef**]
44. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [**Google Scholar**] [**CrossRef**]
45. Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: A deep learning method to predict protein-ligand binding affinity. *Brief. Bioinform.* **2021**, *22*, bbab072. [**Google Scholar**] [**CrossRef**]