Ligand-Target Affinity Prediction and Scaffold Optimisation: Deep Learning Approaches to Accelerated Small-Molecule Drug Discovery

Dr. Minh Nguyen, Professor of Information Technology, Hanoi University of Science and Technology, Vietnam

1. Introduction to AI in Drug Discovery

AI transformation in the drug discovery area is considered to be one of the most exciting and promising trends in the pharmaceutical industry. Without a doubt, the strategies they use should work. Accumulating experimental data and preclinical research on single drug candidates may take an enormous amount of time and resources, without any promise of the expected benefit. This fact stimulated the pharmaceutical industry to target the use of modern technologies to be more efficient and effective in the discovery and design of new medicines. After a brief period of skepticism, AI technologies, which have been proven for decades to accelerate innovation processes in many technical and fundamental science areas, have become widely acknowledged as the new path to generate returns on R&D investments in the drug development area. Numerous examples in recent years have provided examples of how these methods can reduce drug discovery timelines and costs by decreasing the need for individual experiments.

Successful case studies accentuate that these methods might significantly improve effects compared to traditional bio and cheminformatic approaches. Traditionally, biologists and chemists use a multitude of cell assays, animal models, and clinical trials to check hypotheses regarding a drug's ability to treat diseases and to suppress side effects. The volume of data that needs to be taken into account to finalize a drug after many years of work can be tremendous. The level of complexity of our treatment responses to a scientifically verified drug makes it close to impossible to predict the level of success of a single experimental snapshot. The AI-driven approach allows benefiting from analyzing various patterns hidden in voluminous data, making informed decisions much earlier during the discovery phase.

1.1. Historical Context and Evolution of AI in Drug Discovery

1.1. Computational/Artificial Intelligence and Drug Discovery Historical Context The first applications of AI and machine learning techniques to the pharmaceutical industry date back to the 1970s. Classical expert systems and rule-based approaches were merged to some degree with early AI themes and fields for their successful utilization in pharmaceutical and biomedical settings. The first successful applications were based mainly on classical AI methods and domain expert approaches, rarely combining chance and fuzzy logic-based methods or computational chemistry. Some techniques emphasized drug development and were more common in the 1990s. AI has seldom been used in the initial stages of drug discovery to support the selection of individual compounds or the choice of the library subset in high-throughput docking or virtual screening enterprises. However, major technological advances and transformations occurred in the 1980s and 1990s, which have since determined the general approach to the development and execution of AI within R&D processes of pharmaceuticals.

AI has settled in various sectors of the industry and academia since then and started to become of increasing interest within a growing scientific community at large. In the following subsections, we will guide the reader through this evolutionary process, highlighting the technological advances that took place in the closely related field of systems biology, since its early definition as an autonomous study object. Some representative examples of historical breakthroughs will serve as the starting point for a comprehensive reconstruction that embraces the state-of-the-art cutting-edge theoretical developments. In the following paragraphs, an account is given of the general processes related to AI adoption within drug discovery and development and the challenges they faced. Subsequent advancements in computational methodologies and AI methods are thoroughly depicted.

2. Machine Learning Models in Drug Discovery

Machine learning models have become pivotal tools for designing individualized strategies. Approximately 7,000 medical conditions were identified to be associated with causal genes, providing potential therapeutic targets and biomarkers. Besides the identification of drug targets, machine learning models have also been applied to estimate drug-drug interactions and predict drug side effects prior to clinical observation. Knowledge mining of clinical evidence also aids the unearthing of human

therapeutic potentials. The large number of potential drugs and their combinatorial regimens impede massive clinical experiments. Various machine learning models have been developed to overcome these obstacles.

There is a clear shift from the concept of blockbuster drugs toward personalized medicine based on large-scale data. Machine learning models are built to learn hidden patterns from large datasets, and hence identify novel functional annotations prospectively. Drug data is usually in a matrix with the values of features providing detailed descriptions of drugs against something else. Machine learning workflows are important for typical therapeutic applications. Pre-processing, feature selection, and data partitioning are important steps in shaping the accuracy of machine learning models.

Machine learning algorithms have been widely applied in drug discovery and design, including supervised and unsupervised models. Different machine learning models possess unique potential and limitations in various applications. As a result, machine learning represents highly complex models with significant challenges existing in model interpretation and market regulation. Machine learning models with relatively high accuracy are not stable in application to complicated biological systems. Despite the limitations, the large amount of experimental results attached with generalizability may enable the development of guidelines for pragmatic clinical trials. In all, machine learning-based novel therapeutic potential mining may provide substantial social benefits in drug discovery.

2.1. Types of Machine Learning Models Used in Drug Discovery

Machine learning models used in AI-driven initiatives within the drug discovery space can be categorized into various types: a) Decision Tree Models; b) Neural Network Models; c) Bayesian Network Model; d) Support Vector Machines; e) Ensemble Methods. Depending on the nature of the data in use and the particular problem that the drug scientist is striving to resolve, the researcher would need to choose the correct model for their research. Researchers are typically using Support Vector Machines and Neural Network models to address their business problems. The upside of neural network models is that they are able to solve problems with large quantities of data more easily and require little data preprocessing. Ensemble methods also have their advantages in the ability to assess a problem statement from a range of perspectives

before generating a weighted average of the results to create an optimum model. Applications for ensemble methods in drug discovery typically involve making predictions about drugs, i.e., which is likely to be the better of two drugs in terms of treatment or safety.

In drug development, machine learning algorithms are used for making in-silico predictions between a pair of data classes. This can be done during first-in-human trials of a new chemical entity to predict its maximal tolerated dose, to discern adverse effects of a new lead molecule, new indications, or as part of ongoing pharmacovigilance. Machine learning algorithms are also used during preclinical studies to evaluate possible cardiotoxicity, hepatotoxicity, and neurotoxicity. Neural network algorithms are used to facilitate metabolomic phenotyping and to accelerate transdermal drug delivery. The growing metrics of the best scientific journals reveal the increasing role and results of using hybrid models of machine learning in the field of drug discovery. Performance metrics used in these models to evaluate their performance include: a) Misclassification rate; b) Area under the ROC curve; c) Sensitivity; d) Specificity.

3. Identifying Novel Drug Candidates

The discovery of novel drug candidates is essential for the continuous development of therapeutic options. Identifying new small molecules that can modulate biological targets is a time- and labor-intensive process that requires the integration of computational and experimental techniques. One of the major efforts in this direction is the concept of virtual screening. Computational approaches are frequently used to discover, prioritize, or re-prioritize large databases of chemical compounds. If the goal is to discover new leads for biological targets, the emphasis is put on the diversity of the compounds selected, though highly optimized analogs can also be sought to bolster programs that are close to delivery.

The first step in the screening process—the docking of a large number of compounds into a structural model of a biological target—is one of the key predictive problems faced in developing new small-molecule drugs. In these approaches, termed molecular docking, molecular libraries are screened against receptor binding sites, and a variety of scoring qualifications are employed to judge how well each compound fits. Although directly optimizing accurate approaches such as molecular mechanics or quantum mechanics would be infeasible, a wide variety of heuristic scores have had considerable

practical success in this role. Several successful case studies have been reported, in which these methodologies resulted in the discovery of small molecule modulators that could be developed into a candidate drug. However, a major challenge in early-stage virtual screening is the ability to predict whether a hit compound—i.e., a physically validated binding molecule—has potential to be developed into a drug, and, if so, what structure-activity relationships are required to achieve selectivity and potency.

Increasingly, machine learning has addressed some of these issues. While applications of computational techniques have had many individual successes in discovering drug candidates or predicting their behavior in cells, the workflow of deadlines united with bioinformaticians and synthetic chemists should make it feasible for a systematic attack on the drug discovery objectives in both research and industry. Crucially, no approach is cut off from the opportunities or constraints of another, and the relationship between the onset and endpoint of drug discovery is, and should remain, continuous.

3.1. Virtual Screening and Molecular Docking

Virtual screening is a widely used approach in the field of computational drug discovery, particularly in hit identification and lead optimization. It is a fast and cost-effective approach for screening large chemical libraries, potentially from millions to billions of compounds, against several targets to prioritize a relatively small pool of virtual hits for further biological evaluation. In the virtual screening workflow, compounds are filtered from large chemical libraries through screen-out filters, following energy-based and structure-based algorithms that simulate how a random compound binds to a biological target and identify how well the ligand binds. These techniques gradually eliminate the unsuitable compounds from millions to thousands, which are deemed acceptable for biological screening, reducing cost and time.

Molecular docking is an in-silico prediction technique used to predict the conformation of a ligand in the binding site of a macromolecular receptor. It evaluates and predicts protein-ligand binding based on the binding orientation and non-covalent intermolecular interaction energies, which are primarily electrostatic, van der Waals, hydrogen bonding, and hydrophobic interactions. Docking becomes an optimistic solution for hit identification and lead optimization. The process of molecular binding can also determine the energy of binding between a ligand and a macromolecule. The energy can be predicted by employing empirical scoring functions in all the docking

software available. It is well accepted in the pharmaceutical industry that a correlation between calculated binding affinities and the real biological activities of the tested compounds significantly aids in the optimization of drug candidates. Docking, together with simulations as part of the full computational package, has been proven to be a critical tool for in-silico ADME and, particularly, for predicting potential off-targets. It is best for use in the early stages of drug discovery. Overall, virtual screening techniques account for a large percentage of the all-in costs required to bring a new drug to market. Successful case studies to date have significantly accelerated the time to drug discovery. However, more and more obstacles exist in practice in drug design. Non-classical docking generally poses difficulties for cross-docking because of the requirement for receptor flexibility. In practice, force-field scoring functions are unrealistic because they ignore the dynamics that carry actual information. Most crucially, simple in-silico predictions rarely perfectly match biological activity. As a result, all the drug targets requiring virtual screens are prioritized.

4. Optimizing Lead Compounds

Bridging the early discovery phase with clinical development, the optimization of lead compounds is critical to ensuring that they are as efficacious, selective, and safe as possible for advancing into clinical trials. This step is part of a repetitive process in which the early lead that goes into the optimization is tested in living systems – either in vitro, ex vivo, or in vivo. Based on the outcomes, modifications are suggested to the molecule that would increase its specific activity in the biological system, or enhance other properties to improve the drug-like characteristics. Next, the new chemical entities are synthesized and tested for their biological properties and other drug-like characteristics, and the process is repeated iteratively. The possible changes to a compound structure can be designed following: 1. Chemical derivatization involving changes on the structure by addition of substituents. 2. Change in the core structure of the ligand. 3. The merging of two or more chemical entities on the structure. The power of computational chemistry and bioinformatics in giving shape to this process is of increasing relevance. Through several computational techniques, the impact of chemical modifications on the behavior and properties output can be easy and rapidly predicted. Designing the most relevant modifications in order to increase the physical/chemical properties and to produce better pharmacodynamics and toxicodynamics effect represents a further relevant aspect. The investigation of compounds structure/design

could be of interest for making several hypotheses referred to its interaction with enzymes, receptors, and polynucleotide and in silico ligand predictions. This is very relevant for antiviral drugs which act on viral enzymes. The introduction of computational techniques and an iterative process improves the probability of designing multi-task compounds which is a relevant and powerful structure activity approach. The advantage of machine learning to predict the activity when we change a structure lies in the possibility of simulating virtually the real response. It gives us the possibility to predict the compound behavior to decrease, avoid, or reduce false positive or negative results. The latter is of importance in cases of reducing risk in the prediction of potential liver side effects/toxicities. In general, a statistical model involves the relationship between data and the design of the molecule. A more complex model can be performed by using a combination of several statistical algorithms. Some examples of statistical algorithms are genetic algorithm, support vector machines, decision trees, etc. The kind of data required for the development of the statistical model may vary based on the retrieval of information by cheminformatics or bioinformatics approach. Many machine learning rules can be developed in order to predict the results to compound modification. Despite the above, the statistical models are relevant tools that can be used for designing de novo molecules or to perform a structural optimization according to the design in order to select the molecules that might be of interest for the chemical synthesis. One real-life example of structural optimization is the approach for HIV PR inhibitors. A large database was used to build up the quantitative structure–activity relationships models: at the end of the process, a new hit compound and its posology were suggested. Further data about preclinical tests are missing. A new bioassay was quantitatively empowered by using models. In this context, many statistical tools have been reported as powerful models to detect potential immuno-allergic members in a set of structurally heterogeneous compounds. By using the characteristics of a molecule as independent variables and its biological activity as a dependent quantitative or qualitative variable, a mathematical representation can be deduced using different programs, capable of classifying a compound as active or inactive. Models are relevant if the biological activity is supposed to be multivariate linear or hyperbolic parameters correlated to independent variables, representing the structure-related descriptors. Thus, models have great capability as decision-making tools to predict the biological activity of compounds starting from their structure.

4.1. Quantitative Structure-Activity Relationship (QSAR) Modeling

4.1. Quantitative Structure-Activity Relationship (QSAR) Modeling Quantitative structure-activity relationship (QSAR) modeling is a pillar in the process of lead optimization (LO). QSAR models predict the activity of compounds by analyzing the relationship between the molecular structure of a series of congener compounds and their measured biological activities. Without any assumption on the mechanism of action, the most potent compound is selected as the lead compound. Once a group of lead compounds with diverse chemical scaffolds is available, a novel mechanism of action may be explored, if needed (or vice versa). Thus, as drug discovery is an empirical science, QSAR narrows the range of experimental efforts, in this case selecting the most active group or structural species. To develop a predictive QSAR model, a five-step process is generally followed: data collection, variable selection, model building, model validation, and application/prediction.

Data collection includes obtaining a set of compounds with their associated biological activity, in most cases represented by some relationship with IC50, EC50, 50% effective dose, etc. Descriptor or variable selection relates to handling various molecular representations to enable enhanced simplicity in model building. Variable selection aims to remove irrelevant, redundant, or noisy variables. Model building concerns choosing which prediction method is appropriate, e.g., linear regression, neural networks, support vector machines, etc. Model validation is broken into internal and external validation, the latter of which is the most precious. An external QSAR model assessment offers a reliable measure of the predictive or generalization ability of the model. Leading to external prediction, QSAR models developed to date can be applied to predict the potency, duration of action, toxicities, molecular targets, absorption, distribution, metabolism, excretion, toxicity, and structural motifs for mostly untested chemicals. Although some limitations have been reported, QSAR modeling is certainly a valuable tool in rational drug design and mechanism of action studies. Such models reduce the timeline for the drug discovery process by three to five years in situations when drug candidate-related issues arise during animal trials.

QSAR modeling has several hidden limitations. The first limitation relates to the number of compounds in the available database that are tested. Rarely is the desired set of compounds as extensive as needed for the predictive modeling to be effective. The

second problem concerns modeling approaches that can lead to model overfitting and predictive interference due to non-modeled randomness or noise present in the training set. Additionally, the reliability of the descriptor or variable selection process in reducing the dimensionality of the molecular representation and subsequently affecting model generalization is troublesome. Furthermore, these well-known limitations can be associated with a lack of biological knowledge in molecular QSAR models to demystify the molecular essence of complex biological processes. Both traditional and modern QSAR approaches are grappling with the limitations of the field. An avenue towards improvement in QSAR applications lies within the development and use of machine learning methods. When integrated with traditional QSAR approaches, machine learning models have the ability to increase the efficacy of modeling and prediction. It has been reported that machine learning models are able to extrapolate biological activities for broader chemical spaces than classical QSAR models. In an extensive review, several illustrations of the integration of machine learning and QSAR approaches, examples where machine learning models are more accurate than their QSAR counterparts, and the domains in which they are applicable are presented.

5. Challenges and Future Directions

The use of AI-driven approaches has gained momentum in drug discovery to create a faster and cheaper way to identify, develop, and trial new drugs. However, there are several challenges that need to be addressed to make AI approaches more widely applicable in drug discovery. Among these, issues with data, data quality, and data bias are significant and can have damaging effects on research and regulatory compliance. At the same time, AI methods are not always easy to understand and explain, and developers often disagree on measures of model performance. Furthermore, there is often a mismatch between model outputs and endpoints pertinent to end users. Finally, it is unclear where AI is being used, and how AI systems can be integrated into current workflows has not been defined. Furthermore, pharmaceutical R&D is evolving to focus on specialized therapies, i.e., therapies that are highly individualized and specific. AI models are not being developed to cater to these highly specialized strategies. While it is problematic to integrate AI systems into highly specialized strategies now, the research landscape may look completely different in a few years, and an AI-integrated process can and should be adapted to these changing approaches. While these are critical issues, there is an opportunity for providing innovative solutions to address these barriers. To

overcome the challenges of data and limited information about patient cohorts, AI can develop models for less popular targets. Few AI and drug developers collaborate in real application settings to determine how best to integrate AI with drug discovery. By providing answers to these essential questions, AI use in drug discovery will broaden, accelerating safe drug development. Additionally, the ethics involved in integrating AI in drug discovery were not addressed; these are issues that the pharmaceutical industry takes seriously and need to be addressed as AI techniques develop.

5.1. Ethical Considerations in AI-Driven Drug Discovery

While the value of AI-driven solutions in drug discovery is clear, there are several ethical considerations that must be addressed to ensure that these techniques are implemented responsibly. Guidelines provide an important starting place, offering best practices and identifying potential pitfalls. For example, they might include guidelines to ensure that the data rights of the patient data used in training AI algorithms are protected. One risk is that AI could be used in such a way that it undermines the clinician-patient relationship and decision-making, abrogating patient confidentiality. Additional risks arise from the proprietary nature of algorithms, and hence, the lack of transparency. In healthcare, this is particularly concerning given the potential for concealed biases in the algorithm that could lead to inequitable treatment. For instance, the use of AI in health provision may inadvertently result in discriminatory practices if the dataset used to train these systems includes prior biases. Additionally, there are concerns that the use of AI could reveal patient attributes that are not necessarily useful in decision-making about health or treatment, reinforcing existing stereotypes, exposing patient information, or creating a sense of unfair access to health provision.

A major domain of AI that has been heavily critiqued for the ethical concerns is machine learning. For these reasons, solutions in which AI is used in medicine call for iterative testing, with a focus on accuracy and fairness in predicting future health and responses to treatments. As such, the resulting decision-making process should offer an explanation taking the AI predictions into account. Regulators have considered the implications of AI applications in the pharmaceutical industry and encourage those developing such technologies to critically evaluate the potential risks as well as other factors influencing public health and safety. They stress that while medicines regulatory approaches may vary, the development of AI and ML algorithms must nevertheless

comply with current regulatory standards, be underpinned by good data quality, and comply with ethical guidelines and legislation. They emphasize a need for transparency around decision-making tools that use AI technologies while also promoting and fostering innovative approaches to enable appropriate evidence generation. Ultimately, the successful integration of AI into drug development will require rigorous compliance, preclinical toxicology and clinical testing, scalable data, and above all, public trust.

6.Future Direction

Our envisioned future direction sees AI in drug discovery evolve and expand in several directions: Technologies. Advancements in other technological domains such as cryo-EM, with accelerating advances, as well as more expansive deep proteomic profiling techniques, are expected to eventually halve scientific inquiries and enrich the pool of available methods for use in AI-driven drug discovery. Combining these methods, we here anticipate the next advances of AI in drug discovery, providing both overarching trends and a more granular view of the upcoming developments. We foresee available technological investment and expanding variously into mature, extended, and newly developing methods. Multi-disciplinarity. Although many AI methods today still focus on chemical compounds, over the next two years, AI and computer science advances could move to the study and computational targeting of biomolecules. Technologies include AI advances in genomics and personalized medicine that, when transferred to drug discovery, will boost the speed and the outcome of target identification and more sophisticated pathways to patient stratification and repurposing. To further push the boundaries of ongoing implementation of AI brought forth through earlier investment, we anticipate a closer and more functional integration and collaboration of life science and computing experts in both academia and industry. We expect the heterogeneity of investments and main technologies to coalesce after the year 2023, marking a period of interest among the most pharmaceutical companies to house and develop multiple methods in-house or in partnership with AI specialists. Role and Outcome of AI. The AI application in drug discovery has entered an exciting phase of proven efficiency and is poised to deliver real feasibility outcomes in the next 2 to 5 years. These include fast proof of concept for promising new drugs, successful repositioning of tired drugs, first-in-kind out-license potential for smaller companies, in addition to first-time target identification and subsequent validation. In the late segment of our predicted timeline, data-driven AI will still hold promise inside big pharma settings, although the risk

around Pharma AI will heavily depend on long-term efficiency, continual learning, and deployment of model novelty. Manufacturing co-optimization and ethical quality and adoption will guide research and innovation.

7. Conclusion

AI-driven approaches are expected to bring far-reaching changes in pharmaceutical research and development in the years to come. Thanks to optimizing a myriad of laboratory techniques and overcoming numerous bottlenecks, future drug discovery and drug repositioning projects that do not take advantage of the opportunities offered by using AI/ML-based tools might become unfeasible. There are also numerous opportunities where AI could become a prerequisite or important factor in selecting the best ways to develop projects. Nonetheless, the integration of these modern methods with drug modeling at the systems level may sometimes pose new challenges. Technological developments are growing at an exponential rate, and algorithms have significantly evolved, but they must be validated under current good practices and the peculiar conditions of development in the creative and chaotic framework of projects in the pharmaceutical industry. At the present stage, AI offers professional tools to automate high-throughput, repetitive, and time-consuming tasks but is not a magical solution. Comprehensive management of all the development aspects of basic science, drug discovery, clinical development, and pharmaceutical commercial activities requires thoughtful decisions in a network composed of interconnected stakeholders. Indeed, an increasing amount of data, together with novel computational approaches, will further transform the discovery of potential drugs and their clinical development. To this end, the development of deep learning and AI technologies has substantially impacted the efficiency and productivity of the overall drug discovery industry from lead identification and optimization through preclinical evaluation and into clinical development. This evolution has dramatically transformed biopharmacological research from a formal bureaucratic and paper-based enterprise into a digital, searchable world. In conclusion, all these aspects together are laying down the foundation of a new multifaceted system, supported by regulatory agencies, that ensures the safety and efficacy of drugs and the right use of medical devices. In the near future, the new framework will remain crucial for assisting novel drug and device development to fulfill medical needs unmet by the current healthcare system. AI has become an integral part of drug discovery, and further advances will continue to expand the possibilities in drug

development, providing more life-saving options to those who need them most. Additionally, increasing interdisciplinary collaboration between AI developers, biologists, chemists, pharmacologists, and clinicians, among other domains, will be essential for future drug discovery success.