Structure-Activity Relationship Inference and Virtual Screening at Scale: The Role of AI-Powered Platforms in Transforming Modern Drug Discovery

Dr. Veronica Murillo, Associate Professor of Computer Science, Tecnológico de Costa Rica (TEC)

1. Introduction to AI in Drug Discovery

Introduction Artificial intelligence (AI) platforms designed for drug discovery have made significant improvements over traditional methodologies. Given the very high attrition rates in the pharmaceutical industry, relying on traditional drug discovery methods has not been overly conducive to innovation. These results are not wholly surprising, given that traditional methods such as high-throughput screening and lead optimization have not significantly changed in decades. AI holds the potential to speed up further research activities if a holistic approach to AI application in drug discovery is pursued. With ever-increasing computational and informational resources, applying AI technologies to the entire drug discovery pipeline, from understanding the mechanistic basis of disease to designing drugs, has been an attractive proposition in the early 2020s. The realization of such a holistic approach has been anticipated by both scientists and investors to revolutionize the field, delivering comprehensive insights into the fundamental basis of biology and accelerating the development of safe and effective new drugs. While this is yet to be realized, AI has already facilitated tremendous efficiency improvements and is a driver setting up new standards in discovering and exploring innovative drug candidates. The marked advantage over traditional approaches has prompted all major pharmaceutical and biotech companies to at least partly invest in AI for drug discovery. This review summarizes key developments in the role of AI-powered drug discovery platforms during 2010-2020.

1.1. Overview of Traditional Drug Discovery Process

The traditional drug discovery process is composed of several systematic and predictable stages, from target identification to the market launch of the drug. Biology and pharmacology provide therapeutic concepts, usually involving the study of

diseases, genes, gene functions, proteins, or cells. Chemists then find a compound for the target protein with suitable characteristics that can help to treat the disease. In the current model, which has remained largely unchanged for decades, each R&D department conducts its own development and works collaboratively with other departments. Each stage of drug development follows a predictable program: safety, pharmacokinetics, toxicology, and pharmacodynamics required by the regulatory authority at the time of filing. Consequently, the drug R&D process is complex, lengthy, and expensive and demands the integration of many therapeutic areas and the use of multifaceted technologies. The integrated approach to drug discovery and development typically involves areas as diverse as cell and molecular biology, in vitro testing, in vivo pharmacology, pharmaceutics, chemistry of medicinal products, physical-chemical science, and movement disorders, while non-technical functions include intellectual property and finance.

An extensive and careful investment is required in both time and money to translate a promising treatment into a new medicine, which is a costly and uncertain process associated with high failure rates and the potential for expensive late-stage surprises. Generally speaking, when drug R&D becomes unexpectedly more challenging and costly, fewer medicines may reach the market when looking at the traditional R&D model from a macro perspective. Projects with promising individual drugs are expected to fail due to safety or lack of efficacy; long timelines prevent portfolios from being refueled efficiently with new innovative projects, thus severely undermining sustainability. We believe that the AI industry plays an important role in identifying hidden correlations and eventually in reducing inefficiencies and unnecessary costs in these R&D models. Empirically based thoughts and decision-making often fail to produce drugs that can succeed in clinical research. The complexity of living organisms is difficult to address solely through empirical approaches. Moreover, following the repeated failure of biological drugs in Parkinson's disease, the world of drug development has entered a new phase of skepticism about the conventional wisdom of science. The possibility of appropriately combining a technology-driven approach in drug discovery and drug development processes serves as a warm encouragement for both the biotech ecosystem and major pharmaceutical enterprises to overcome these fundamental limits.

1.2. Emergence and Evolution of AI in Drug Discovery

The adoption of AI in drug discovery presents the latest turn in technological and methodological transitions that have spanned nearly a century. After an initial "golden age" prior to the 1960s, where systematic methods were developed for toxicity prediction and chemical synthesis, the following decade would see chemodynamics analysis with an emphasis on quantification. Machine learning and data analytics began to complement model-based approaches for drug development, being used for early simulations of personalized medicine. The combination of rational drug design and machine learning has been attributed to fully enable model-based systems medicine, from DILI prediction to model-based therapies. There were several milestones that facilitated and directed the adoption of AI.

First and foremost, AI technologies have become mainstream, partly due to breakthroughs in computational power and the rise of big data, as well as the development of agile methodologies. This led to the development of several software platforms that have since been marketed and employed. AI has offered improvements despite deficiencies, which even affected certain technologies. These factors led to AI being integrated into machine learning nodes for the training and validation of QSP models, pharmacokinetics models, and clinical outcome models. The second milestone follows the rise of big data, which reached terabyte scale in genomics research before most others; clinically, it had reached gigabyte scale in imaging diagnostics. With terabyte-scale proteomics and genomics data, it became time-efficient to build AI-based disease and patient classifiers that stood to change drug discovery research paradigms. One research study in AI in this field was exemplified with a study that combined qualitative physics with supervised machine learning in order to obtain efficient biomarkers, as well as disease and patient classifiers for certain patients. This research was the starting point for the development of a then unique AI-based drug discovery platform.

2. Machine Learning Algorithms in Drug Discovery

Supervised, unsupervised, and reinforcement learning are the main categories of machine learning algorithms that have been instrumental in the field of biology and biomedicine. Supervised learning algorithms are used in drug discovery in a variety of ways, ranging from drug target prediction to predictive toxicity models. Similarly,

unsupervised learning can group interactions of potential drug candidates and their targets or group patients based on how a drug is expected to interact with their body. However, the most important and critical application of ML in drug discovery has been in reducing the time it takes to screen millions of potential drug candidates by using public gene expression data generated across different diseases and cell models to reduce the hit compounds down to several dozen candidates for further experimentation.

Supervised and unsupervised learning algorithms together use several types of MLs to power a new generation of platforms that reduce the number of drug candidates that hit a target from tens of thousands of chemical structures to only a handful of molecules by incentivizing the prediction of previously validated drug target interactions. Additionally, due to the practical possibilities of running experiments, MLs can integrate diverse pharmaceutical datasets of selective and non-selective drugs, antibody-drug conjugates, healthy and disease-related genes, genetic and pharmacologic perturbations, and chemical and genetic genomics profiles into a consistent multi-molecule knowledge base that does not only profile the binding specificity of molecular bioactive molecules, but also therapeutics and interventional drugs. With this new class of platforms, called AI-powered drug discovery platforms, the number of diseases that these MLs can screen is only bound by the public datasets available that contain at least two omic frames of information as inputs. Once trained, the cost of running new experiments needed to calculate a new drug-disease score for any new therapeutic indication that the platform was not trained on is very small, limited only by the throughput of the experiment. With this adaptability, promising drugs can be brought to market more quickly than traditional routes of development.

2.1. Types of Machine Learning Algorithms Used

Machine learning is increasingly being used in different stages of drug discovery and can roughly be divided into three most common algorithm types: the first category includes rule-based models, such as decision trees, that explain their decision-making process and have the advantage of being easily interpretable. The second category includes the so-called non-rule-based models, which are less transparent but may yield differences between rules that were recently classified as active or inactive according to the Tox21 challenge.

In addition, however, other machine learning models are being used, such as the more biologically inspired neural networks and support vector machines, for performance reasons exemplified by the work to refine the default structures of HQSAR models, or other robust methods such as support vector machines. A meta-analysis of about 100 quantitative structure-activity relationship and/or quantitative structure-property relationship models from 20 independent toxicity studies on 12 databases illustrates the capabilities of 11 different machine learning algorithms. Furthermore, there is a trend to combine the best of both worlds by using so-called expert systems, which may, for example, use decision trees to predict in vitro-in vivo ADME relations and analyze structural alerts. In contrast, non-experts might benefit from tools that can make use of thousands of features, pattern-recognition capabilities, and other advantages.

2.2. Applications of Machine Learning in Drug Discovery

The present section will provide an overview of modern attempts to integrate machine learning in different stages of drug discovery as a practical application of the previously described algorithms. For each stage of AI implementation, examples of developed technologies will be presented. The highest number of examples was found for compound screening, lead optimization, and omics data-based technologies, such as pharmacological profiling.

Although current drug discovery and development processes are relatively inefficient and costly, several diverse AI/ML-based interventions have revolutionized DDD strategies. Predictive analytics are expected to aid and lead to feasible healthcare administrative activities in addressing issues of clinical trial failure and adverse events. Conventional drug commercialization has proven to be less innovative and has seen drug approvals with restricted health utilities. It is expected that these machines will change and improve the trajectories of drugs from bench to bedside. It will usher in a new phase of drug production by reducing the need for several arduous discovery steps. Conclusively, implementers now firmly believe that the translation of current AI advances into drug discovery, coupled with advances in bioinformatics, will provide stronger and more efficient treatments targeting different subgroups of diseases.

3. Challenges and Opportunities in AI-Driven Drug Discovery

AI-driven drug discovery faces numerous challenges, some of which are directly linked to the data it is trained on and operates with. Not all data is equally useful for learning

about mechanisms of disease pathobiology or the biological activity of small molecules. To benefit the development and validation of sub-models within AI-driven technologies and to build general predictive systems, diverse data from numerous sources would be advantageous to prevent models from learning spurious patterns. The increased collection of multi-omics datasets is anticipated to help address this data quality concern.

Regulatory and data privacy issues are often the biggest arguments against sharing data and models, as the healthcare sector is under-protected for a variety of technical reasons and by processes designed for patient protection above all. AI is making more decisions in the diagnosis and treatment of patients in hospitals. Some drug discovery economists reason that the exact timing of the fruits of their AI investments becoming evident will depend on structure leads, such as improved algorithms, data types, data quality, and model explainability. Concerted action from all stakeholders, well-managed negative headlines could slow business contracts, but generally, AI represents an extremely powerful and transformative set of tools for drug discovery. Regulatory pressures continue to influence the growth of collaborative research around AI-driven drug discovery, and in the future may make direct access to clinical data more difficult, adding to the advantages of being able to do research with clients and data available across the research communities and industry.

3.1. Data Quality and Quantity

Data quality and quantity are main influencing factors for data-driven AI applications in the pharmaceutical industry. Data quantity sets limitations on the number of input variants, while data quality can improve the prediction potential of AI models. In AI, the quantity of available data can also enhance the predictive power of ML models and the performance of transfer learning techniques. However, quantity alone cannot ensure good generalization to a variety of use case scenarios. For practical applications in actual settings or deployment of models in remote geographical areas, AI models should generalize well, which is a combination of high-quality data sources coupled with diverse data. These factors alone will provide the necessary information needed for enhancing the generalization of AI models. Real-world datasets are sometimes incomplete or biased, which may lead to poor prediction accuracy of the ML models. For drug discovery applications, the use of low-quality datasets generated inconclusive

findings due to using incomplete or biased data. We hypothesize that this problem can be solved by synthesizing necessary information in real-world databases, thus improving the chances of generating high-quality datasets for good prediction accuracy.

Data quantity and diversity also matter. The actual disease or parasite data should be diverse enough to cover all areas of interest, especially in the context of personalized medicine. Here, one cannot study 100% of infected individuals, but instead can study a situation where a subset of infections arises for a certain geographical location, age group, or gender. We can then model the situation using diverse infection datasets and synthesize data or models that can be accurate enough for the sub-population if no record exists. This theory also demands that the sub-populations synthetically generated should be dependent on the real infection data to avoid the generation of unrealistic data. For most drug discovery and AI applications, the availability and use of data, in the form of databases or datasets, are imperative. Academic research sharing data formats and methodologies involving several diversified researchers stimulates the use of large data sources in drug discovery for data-driven computational applications. Moreover, academia and the pharmaceutical industry can collaborate for data sharing to improve AI methodologies in real-world applications. However, the use of authentic data from pharmaceutical research is still more encouraged because it is closer to the ground truth. Synthetic data may further be useful when data freedom, privacy, and confidentiality issues exist, where real-world-based data sources may no longer be available to the public or for research use.

3.2. Regulatory and Ethical Considerations

Regulatory and ethical considerations: Incorporating AI into the drug discovery process involves utilizing big data. However, the big data used for AI model development needs to comply with different existing regulatory frameworks separately. Regardless of the specific big data and the applicable regulations, the regulatory, patient safety, and drug efficacy considerations are always first. Most of these regulatory policies for new drugs are the same in the sense that the therapeutic effect should be greater than or equal to their side effects. The efficacy and safety data in the submission dossier must be high quality. Drug development and discovery are inherently regulated fields. The development and improvement of analytical tools represent a normative shift, an improvement within the field that requires transparency and compliance with ethical

norms. Emulating AI decision-making models is inevitable and should be consequential. Transparent AI and other initiatives have worked on soft law development as governing principles, guidelines, and an ethical matrix on the ethics of AI. For AI to be used responsibly for drug development and discovery, guidelines may need to be prepared. For the fallback option, the requirement and difference lie within these-to-remain avenues. Thus, unused medicines can keep exploring new types of exhibition and uses. Off-label uses are originally approved medical products that are used other than their approval. Harmonizing global rules will result in a worldwide standard that should avoid regulatory arbitrage, whereby an AI model is validated in one country and therefore approved quickly in others. While a unified approach would have implications for both the U.S. and internationally, the potential benefits on public confidence, trust, investment, competitiveness, patient safety, and a global standard that could modernize and streamline the regulatory review of AI-driven drug development are very apparent.

4. Case Studies and Success Stories in AI-Driven Drug Discovery

We provide several examples of drug discovery success stories from applying AI. Each example illuminates a different part of the drug discovery process and showcases the innovative methodologies being used to do cutting-edge work. Projects span different disease areas and assets, including novel small and large molecules, gene therapy, and RNA projects at various stages, including idea generation, preclinical phase, clinical phase, and regulatory approval stage. These research collaborations explicitly rely on complementary skill sets and input from interdisciplinary researchers.

These examples highlight opportunities and ways we can increase confidence in AI technologies for drug discovery. When possible, we discuss potential scientific or methodological lessons and show how these can inspire future work. The decision to include other publications was based on their level of innovation, high potential impact, or informative and understandable explanations that could increase reader access to new developments. When we do not mention a specific piece of technology discussed in the paper, the decision was based on high reader interest in the application of AI to the research question, rather than on the specific details of the advanced technology. Note that some of these examples are historic, and advancing the state of the art may already surpass these early attempts.

5. Future Directions and Trends in AI-Powered Drug Discovery

Beyond improvements in the performance of AI techniques, advancements in technological capabilities could further enable AI-driven drug discovery. Developments in quantum computing have the potential to revolutionize machine learning-based drug discovery. In the future, AI-driven platforms are expected to bring more effective drugs to market by enhancing the application of personalized medicine and the generation of real-world evidence for pharma. AI models anticipate the variation in individual responses to treatment but also narrow the number of experimental groups. Evidently, AI in recruiting patients and real-world data collection has the potential to better approve drugs on reduced clinical trial datasets—an aspect critical to accelerating innovation in drug discovery.

The collection and analysis of big data in real-world settings allow for the development of digital twins—mirroring the characteristics of individual patients on a computer. In the future, AI is expected to drive the rise of decentralized clinical trials by facilitating the delivery of products at home, accelerating recruitment, and overall trial management. Ongoing research activities will likely shed light on the potential future performance configurations harnessing AI. These focus areas include but are not limited to the prediction of unforeseen molecular targets, multi-drug combinatorial interactions, beyond rule-based programs, three-dimensional structures of drug-target complexes, chemical reaction prediction, multi-scale models of drug distribution in tumors, adaptation of the short trajectory of AI to the long trajectory of clinical development and industrial manufacturing. Collaborative technology and tech companies begin to pave the way for innovation in drug discovery and development. Clearly, AI is a fuel rod for the future of health care, and the potential systems effect should encourage deeper and accelerated transformative investment in the movement. With ongoing refinements in AI and novel processes, the future AI-driven drug discovery approaches can only refine the areas already covered and in response to the research trends outlined.

6. Conclusion

AI-based technologies constitute a new era in medicine and research, shifting the incumbents in the biomedical fields on many levels. Drug discovery offers an early glimpse of AI's potential: while traditional methods have proven slow, laborious, and with extremely varied rates of success, the first applications of AI have shown

dramatically improving those metrics. In this essay, we have reviewed the operational methods of AI-driven drug discovery platforms, highlighting ways to make a traditional, effective, medicinal discovery process faster and applying it in an unbiased manner, open to new hypotheses.

It is clear that tools based on machine and deep learning, as well as reinforcement learning, have improved the discovery timelines, raised the rates of pharmacologically active molecules from virtual screenings, enlarged the chemical space of druggable compounds, and eliminated the biases present in many discovery methods. Still, methodological technological bottlenecks are challenging to address, especially in ensuring data quality. Perhaps more challenging are the regulatory hurdles ahead. The goal of a complete drug discovery-to-market journey is to have a truly profitable drug reaching patients. For that, the lending of the biopharmer conglomerates, investors, regulatory bodies, clinical practitioners, and patients is required. From an R&D point of view, each of them holds different stakes and they must be ensured that their specific concerns are addressed. Hence, the success of the drug discovery field will depend on them engaging in the technological efforts every platform and pharmaceutical company are continually developing. Biomedical innovation as a whole relies on the cooperation between all stakeholders, from data scientists to patients.

Looking into the future, a newer, more collaborative view for the full development cycle needs to be ensued. A broader outlook, attracting and embracing the scientific and the entrepreneurial thinker, needs to be promoted. Interdisciplinary units need to encompass scientists and doctors, bioinformaticians and coders in order to design and create more ambitious R&D tasks by leveraging the advancements from the entire ICT ecosystem. It is through these ground-up efforts that a newer, more connected ecosystem can be created. In such a holistic and interconnected system, the adoption of more concentrated, cognitive algorithms and research tools is purely determinant in the medical imposition, care and service. Such a phase is of course fraught with ethical implications, and the latter need to be skilfully addressed to allow AI to continue to build a healthier and more efficient world.