AI-Driven Predictive Modeling for Drug Target Identification

Jose Antonio Talamantes
Farmingdale, United States of America
talamantesja@gmail.com

***Abstract*— This paper reviews the current applications and future potential of AI in drug discovery, focusing on predictive modeling, chemical synthesis, drug design, drug repurposing, and the role of virtual screening. The discussion includes an evaluation of the challenges and opportunities presented by AI technologies, highlighting the transformative impact they may have on the pharmaceutical industry.**

Introduction

Artificial Intelligence (AI) is starting to take center stage in the medical world as a transformative tool. AI can offer solutions that were previously unimagined by humans. Not because humans are incapable of creating these solutions, but the natural limitations presented by time AI can augment us to overcome them. AI can rapidly compute and evaluate numerous potential solutions; this allows AI to offer solutions that traditional methods were met with challenge when faced with. This paper is a comprehensive review of AI-driven predictive modeling used for drug target identification. We will cover methodologies and highlight the strengths and shortcomings of each. The paper concludes by examining future research directions, focusing on how AI is accelerating the identification and development of innovative drugs.

I. AI IN PHARMA

AI entering the pharma field is another tool being added to try and help bring down the cost associated with developing new drugs. "The discovery and development of a new drug is an extremely long, costly, challenging, and inefficient process that takes an average of 10–15 years" [5]. Heightening the stakes is the cost which can balloon to well over US $2 billion dollars. Even then out of 10 therapeutic molecules only one will pass the Phase II clinical trials and receive regulatory approval. Failure to deliver a viable drug can result in massive financial burden on a company. One of the primary challenges in drug discovery is the sheer vastness of the chemical space, the set of all possible molecular and solid-state compounds, is almost unimaginable. According to (Sanchez-Lengeling and Aspuru-Guzik) the chemical space project using at most 17 heavy atoms has mapped out 166.4 billion molecules [1]. When considering small molecules though, there is an estimated 1060 structures. AI thus becomes a natural outlet to try and solve this issue. There are several hurdles that must be overcome though, "challenges that prevent full-fledged adoption of AI in the pharmaceutical industry include the lack of skilled personnel to operate AI-based platforms, limited budget for small organizations, apprehension of replacing humans leading to job loss, skepticism about the data generated by AI, and the black box phenomenon" [2]. As more methods are used and more AI professionals enter the field, a better understanding of how these algorithms generate solutions may alleviate some of these concerns. Currently, a variety of AI-driven methods are being explored to innovate and expedite the drug development process.

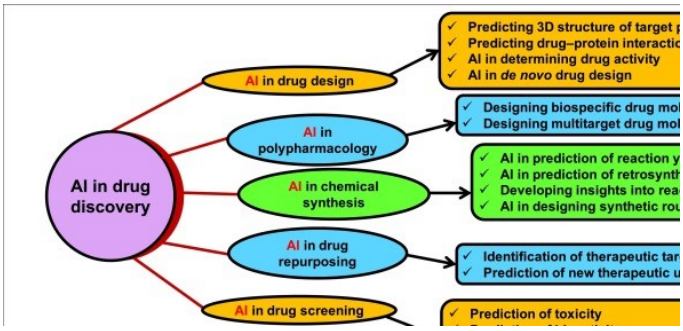

Figure 1
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577280/

II. AI IN CHEMICAL SYNTHESIS

AI can predict the reactive yield of compounds before they are physically synthesized. Given that the virtual chemical space is so big, companies are constantly looking for faster and cheaper ways to navigate the arena. "The efficacy of drug molecules depends on their affinity for the target protein or receptor. Drug molecules that do not show any interaction or affinity towards the targeted protein will not be able to deliver the therapeutic response. In some instances, it might also be possible that developed drug molecules interact with unintended proteins or receptors, leading to toxicity" [2]. Therefore, drug target binding affinity (DTBA) is vitally important to drug target interaction predicting. "AI-based methods can measure the binding affinity of a drug by considering either the features or similarities of the drug and its target" [2]. To illustrate the efficacy of these AI methods, let's consider the Tox21 Data Challenge organized by the National Institutes of Health, Environmental Protection Agency (EPA), and US Food and Drug Administration (FDA). The challenge was centered around evaluating several computational techniques forecasting the toxicity of 12,707 different environmental compounds and drugs. A contestant, DeepTox a machine learning algorithm stood out by successfully identifying static and dynamic features, thereby accurately predicting molecular toxicity.

PrOCTOR is another method used but is trained on the Random Forest(RF) model. PrOCTOR assesses drug likeliness properties, molecular features, target-based features, and the properties of the targeted proteins. A PrOCTOR score is created from these, the score the PrOCTOR score, helps predict the likelihood of a drug failing clinical trials due to toxicity. The scores are also able

to be applied to already FDA-approved drugs, this can be useful in finding drugs that will have adverse effects. This is possible because "AI techniques can analyze large-scale biomedical data to identify existing drugs that may have therapeutic potential for different diseases." By repurposing approved drugs for new novel uses, AI accelerates the drug discovery process and significantly reduces costs.

III. AI IN DRUG DESIGN

Another way AI is being used in drug design is in the precise prediction of molecular structure. Crucial in ensuring that the proper structure needs to be available to connect to the proper receptor. Proper molecular configuration is needed to elicit the desired chemical reaction, it is also needed to ensure that a drug correctly attaches to the appropriate protein receptor. "Numerous proteins are involved in the development of the disease and, in some cases, they are overexpressed. Hence, for selective targeting of disease, it is vital to predict the structure of the target protein to design the drug molecule" [2]. AI can assist in structure-based drug discovery by predicting the three-dimensional structure of proteins. This is pivotal for anticipating a drug's efficacy and safety before synthesis begins. By knowing the needed structure, we can anticipate the effect of a drug on the target. One of the tools being used in this space is AlphaFold, an AI tool that uses Deep Neural Networks (DNNs) has been able to do just that. AlphaFold can "analyze the distance between the adjacent amino acids and the corresponding angles of the peptide bonds to predict the 3D target protein structure" [2]. With such advanced capabilities, AI not only streamlines the drug design process but also enhances the predictability of drug interactions, paving the way for safer and more effective therapeutic solutions.

IV. AI IN DRUG REPURPOSING

AI's capability naturally lends itself to drug repurposing. Using the 3D structures of the drugs and the target proteins, AI can efficiently explore new therapeutic uses for already approved drugs. This accelerates drug discovery and reduces cost. This is possible because "AI techniques can analyze large-scale biomedical data to identify existing drugs that may have therapeutic potential for different diseases" [3]. AI considers "similarity-based interaction(s), the similarity between drug and target is considered, and it is assumed that similar drugs will interact with the same targets" [2]. This methodical approach enables a more precise targeted exploration of new applications for existing drugs. This highlights the impact AI can have on the field of pharmaceuticals.

V. VIRTUAL SCREENING AND THE FUTURE OF AI IN PHARMA

All of these applications together really fall under Virtual Screening. However, for these methods to reach their full potential, improvements in inter-method communication and understanding are necessary. Currently, each method faces specific limitations. Some "models are some way from the predictions of complex biological properties, such as the efficacy and adverse effects of compounds" [1]. For others it is "not uncommon for docking simulations to result in the discovery of inactive molecules" [3]. "The entire success of AI depends on the availability of a substantial amount of data because these data are used for the subsequent training provided to the system" [3]. Prior to 2004 comprehensive data was scarce and limited the capabilities of AI in drug development The release of major databases PubChem and ZINC in 2004, followed by DrugBank and ChEMBL in 2006 and 2008 respectively, was a significant turning point. These databases have not only gave access to the training data AI systems needed but have also created a push of studies focusing on AI's role in advancing drug discovery

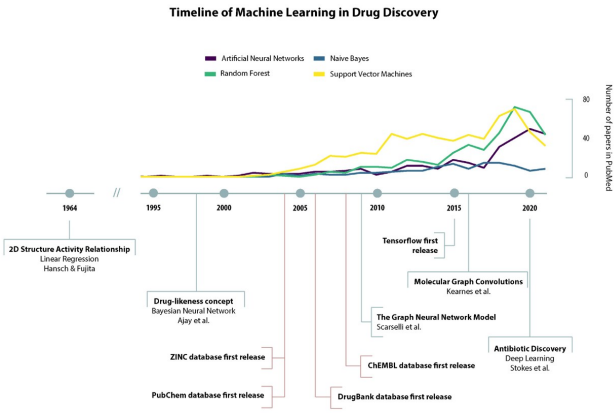

Figure 2
https://www.sciencedirect.com/science/article/pii/S2001037021003421#s0180

In this graph we can only a roughly combine 210 papers were published in PubMed, a database developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S National Library of Medicine (NLM), located in the National Institutes of Health (NIH). "PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally. The PubMed database contains more than 37 million citations and abstracts of biomedical literature" [6].

VI. NATURAL LANGUAGE PROCESSING IN DRUG DISCOVERY

With comparatively so little research having been done in AI-driven drug discovery, it's understandable why communication between the various branches remains limited. However, communication is a cornerstone of progress in any field. Just as we use English to share ideas and understand concepts, like a simple knock-knock joke. We can also understand this code because this is another language we've learned

```
import random
import os

number = random.randint(1,10)

guess = input("Silly game! Guess number between
guess = int(guess)

if guess == number:
    print("You Won!")
```

If we combine the concept of language with chemistry, we get SMILES (Simplified Molecular Input Line Entry System). SMILES "translate(s) a chemical's three-dimensional structure into a string of symbols that is easily understood by computer software" [7]. SMILES is different from Condensed Structural Formula, the method we are most used to seeing, in that SMILES is language designed specifically for describing the structure of a chemical species using short ASCII strings. With a language in place, we can turn to using a technology we have spent several decades developing, Natural Language Processing. Natural Language Processing works by predicting the next letter or word that a user wants based off what has been input already. This same concept can be applied to chemistry. "To model molecules instead of language, we simply swap words or letters with atoms, or, more practically, characters in the SMILES alphabet, which form a (formal) language" [8]. "The physics and chemistry of these molecules are governed by quantum mechanics, which can be solved via the Schrödinger equation" [1]. Knowing the rules of the language allows us to write new molecular stories. The way the model works is a probability is calculated from a seed denoted by s1, as new characters are added the model keeps updating, adhering to the established chemical language rules.

$$P_\theta(S) = P_\theta(s_1) \cdot \prod_{t=2}^{T} P_\theta(s_t | s_{t-1}, \ldots, s_1)$$

Figure 3 https://pubs.acs.org/doi/10.1021/acscentsci.7b00512

This method can be invaluable to chemist who currently must comb thorough a large chemical space looking for molecules that are active and will properly affect the target protein. "Active means for example that a molecule binds to a biomolecule, which causes an effect in the living organism, or inhibits replication of bacteria" [8].

Conclusion

In conclusion I believe integrating Natural Language Processing (NLP) into the drug discovery process can significantly enhance efficiency and reduce the costs associated with new drug development. If NLP is used to simulate through millions of structures research can quickly identify compounds that are viable, not just something that in theory but can also be practically synthesized. Using NLP to narrow the chemical space to compounds with the highest potential researchers would be able to focus their efforts more quickly. From there other methods can be used to assess the efficacy of the compound and verify its structural suitability for binding with the intended targets. Using an assembly line like approach would ensure the compounds not only exist in theory but are also functional and effective in practice. Ultimately, NLP serves as a crucial tool to overcome the initial hurdle of selecting the most promising candidates from an overwhelming array of possibilities.

REFERENCES

[1] Sanchez-Lengeling, Benjamin, and Alán Aspuru-Guzik. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering." Science, vol. 361, no. 6400, July 2018, pp. 360–65, https://doi.org/10.1126/science.aat2663.

[2] Paul, Debleena, et al. "Artificial Intelligence in Drug Discovery and Development." Drug Discovery Today, vol. 26, no. 1, Oct. 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577280/

[3] Vora, Lalitkumar K, et al. "Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design." Pharmaceutics, vol. 15, no. 7, 10 July 2023, pp.1916-1916, www.ncbi.nlm.nih.gov/pmc/articles/PMC10385763/, https://doi.org/10.3390/pharmaceutics15071916.

[4] Author links open overlay panelPaula Carracedo-Reboredo a 1, a, 1, b, c, d, e, Highlights•Machine Learning in drug discovery has greatly benefited the pharmaceutical industry.•Application of machine algorithms must entail a robust design in real clinical tasks.•Trending machine learning algorithms in drug design: NB, & AbstractDrug discovery aims at finding new compounds with specific chemical properties for the treatment of diseases. In the last years. (2021, August 12). A review on machine learning approaches and trends in Drug Discovery. Computational and Structural Biotechnology Journal. https://www.sciencedirect.com/science/article/pii/S2001037021003421#s0180

[5] Lavecchia, A. (2019). Deep learning in drug discovery: opportunities, challenges and future prospects. Drug Discovery Today, 24(10), 2017–2032. https://doi.org/10.1016/j.drudis.2019.07.006

[6] PubMed. "PubMed Overview." PubMed, National Library of Medicine, 15 Aug. 2023, https://pubmed.ncbi.nlm.nih.gov/about/

[7] Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001 Appendix F. SMILES Notation Tutorial

[8] Segler, Marwin H. S., et al. "Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks." ACS Central Science, vol. 4, no. 1, 28 Dec. 2017, pp. 120–131, https://doi.org/10.1021/acscentsci.7b00512.