Generative Molecular Modelling for Orphan Disease Therapeutics: Deep Learning Approaches to Target-Specific Drug Discovery

Dr. Sudarshan Bhattacharyya, Professor of Computer Science, Indian Institute of Technology Kharagpur (IIT Kharagpur)

1. Introduction

Rare diseases are diseases that affect a small number of people when compared to the general population. In the United States, a rare disease is defined as a disease or condition that affects fewer than 200,000 individuals. Seven thousand rare diseases are known and are estimated to affect 25 to 30 million Americans. Of those seven thousand rare diseases, only approximately 5 percent have treatments. However, the barriers to developing new therapeutics are multifaceted and include economic, regulatory, and scientific challenges. In this work, we address a scientific challenge of identifying chemical compounds that bind to a target. We utilize a deep learning-based method and investigate its performance as a drug design platform for targets involved in rare diseases. The method takes as input training data that consist of 3D structures of small molecules and 3D structures of protein targets, along with their affinity to bind as measured by various metrics. It outputs a deep learning model that can predict the likelihood that a given small molecule and a target will bind.

Our goal is to explore and leverage the potential of the method to be effective in candidate therapeutic compound prediction for rare disease targets by comparing its performance across and within disease types based on the dimensionality of the input representation for 3D structure. Herein, we use learned models that were previously provided with training data concerned with either three-dimensional ligand-based molecular similarity or protein sequence/structure similarity models. We observe that the method has higher performance for targets that had available binding target training data and that our version of the method with 3D structure input also outperforms the sequence-based model for targets with a binding target. By utilizing the method to learn

from and make predictions across different rare diseases, we identify drug compounds that are cross-indicated and have the potential to treat multiple rare diseases of unmet therapeutic need.

1.1. Background and Significance

Genetic and protein-level investigations have been the subject of several studies in the case of many rare diseases, and the usually available gene mutation data makes these conditions especially good targets for drug repurposing. Nevertheless, most rare diseases are still orphan, i.e., they have no known beneficial treatments. In many cases, third-party medication becomes necessary to alleviate the symptoms. It is also common that terminally ill individuals consume several drugs in parallel. It is not rare that a rare disease is investigated in the special topic of a physician's or a clinical pharmacologist's study. We cannot expect that the condition and therapeutic principles of these trainees are preserved, and the invaluable knowledge is exploited at the best level easily. AI-enhanced drug design can therefore significantly enhance the development of new therapeutic substances. It has the potential to be a cheap, fast, and easily scalable alternative to other therapeutic development projects. Since repositioning of an existing drug is cheaper and faster than screening newly developed substances in vitro or with fewer ethical requirements, many experts say that drug repositioning can become a leaping frog at the moment.

There are mainly two reasons why most rare diseases have no current remedy. (i) Private companies see that the market is too small, which is not attractive enough for investment. It is, however, proven that in the last decade, decisions of committed local pharmaceutical industries changed lives and increased the quality of life for many cases. Approximately 3,500 different diseases affect 300 million patients, and I am not counting the half million new diagnosis cases every year. I think that there is not a statistical proof that all together with the momentary help for the quality of life of the affected is attractive for any high-level official decision and pushes for more preventive bio-medicine and pharmaceutical research there. (ii) Mature pharma companies consider it strategically reasonable to buy companies in the clinical trials phase because of the new technology or new working mechanisms that are directly shown in humans. However, throughout the last decade, many rare diseases were identified, and various causative agents were specified by significant health care financing initiatives. I am of the opinion

that many rare diseases whose discovery condition came with exceptional success were described partially. They had enough active, fervent followers who could attract relatively much involvement. During parallel results of other studies, similarities in the particular characteristics were formulated, and somewhat different clinical pictures, with the common knowledge help of time, decreased the number of rare diseases. To summarize, with the passage of time, due to recent developments, the number of unidentified diseases decreases among rare diseases, so more and more orphaned people become involved. Therefore, the question arises: How to support the huge amount of emerging technical human curiosity?

1.2. Scope and Objectives

Although the technical abilities of structure-based drug design have gradually become practical, the scarcity of public information on rare diseases hinders the bottleneck of drug discovery for these conditions. The present study exhibited the feasibility of overcoming this obstacle by exploiting artificial intelligence-enhanced computational techniques. Accordingly, case studies are exclusively presented for approved drugs of orphan indications. SETCOM, as an advanced deep-learning-based 3D drug design approach, was adopted. Our analyses further exemplify the versatility and applicability of in-house techniques and present collaborative techniques for the response to urgent public health crises, which can be customized for any declared disease other than the case. Since the breakthrough of molecular modeling technology, structure-based drug design has emerged as a complementary bridge to traditional serendipitous or phenotypic screening processes in pharmaceutical pipelines. Rational drug design techniques have endowed drug hunters with computational insights on the molecular recognition between drugs or drug derivatives and the drug targets suspected to be disease-specific. The association of known medications with recruited orphan indications or rare disease targets can expand the scope and broaden the utility of identified drug candidates. It also takes advantage of the safety and pharmacokinetics information for the conversion of known drugs from bench to bedside.

2. Understanding Rare Diseases

The first, and arguably the most significant challenge in tackling diseases with small numbers of validated targets, is the limited understanding of the diseases in question. Many, if not most, genetic mutations linked to CDGs are of unknown function. Since

about one-sixth of human genes impact glycan assembly, there is a vast candidate space that likely contains new CDG genes. Even for known disease genes, the precise impact of every pathogenic mutation decrypted in a clinically observed instance is often unknown, and the already daunting goal of decrypting the genome becomes further complicated. This lack of understanding is common in rare diseases generally.

Part of this challenge, beyond general biological complexity, falls directly to the relatively small CDG community and bodies of scientific knowledge that have been accumulated. Researchers and clinicians focused on CDG-related biology, even when examining specific CDG cases, may be interacting with only one or a few specific gene products directly relevant to their research. However, CDG research as a field cannot be isolated in a bubble: the glycosylation pathway intersects with many other pathways, and the consequences of disparate mutations may not be understood without broader study. Such study may be difficult, particularly where it may require non-standard and potentially cross-disciplinary science. AI, however, excels at drawing connections between disconnected fields and, from this broader perspective, is foundational for meaningful research on not just all diseases but also on general glycosylation mechanisms and signal flow understanding.

2.1. Definition and Prevalence

Rare diseases are diseases that affect only a small percentage of the population. Few scientists have attempted to define such diseases. A disease should be considered rare when it affects less than 1 in 2,000 citizens. A rare disease, also known as an "orphan condition," is defined as a disease or condition with a prevalence of fewer than 200,000 affected individuals, or a prevalence of more than 200,000 affected individuals if there is no reasonable expectation that the cost of developing the drug will be recovered from sales. The prevalence is equal to or less than 1 in 2,000, but the definition changes between different countries.

Currently, between 6,000 and 8,000 rare diseases have been identified; 72% are genetic, and 70% of them appear in childhood. Polygenic rare diseases also exist, but the vast majority are caused by the mutation of a single gene. Approximately 30 million to 40 million individuals have at least one rare disease, totaling 64 million. Eighty percent of the rare diseases are caused by malfunctioning proteins. A drug can also be considered an orphan if it is going to be used for a common disease, but it is only being developed

to treat fewer than 200,000 individuals, or more than 200,000 individuals if there is no reasonable expectation that the cost of developing the drug will be recovered from sales.

2.2. Challenges in Drug Development for Rare Diseases

The long and arduous journey of drug development begins with the recognition of an ailment as a disease and the collection of its clinical data. Only after a molecule implicated in or causative of the disease is found can drug design commence. Successful development may take more than a decade and cost more than $3 billion to bear a single drug in the US. Individuals with rare diseases experience a longer diagnostic odyssey than patients with later-onset diseases. It may be assessed during preclinical and clinical testing because of the scarcity of preclinical tools such as models of the disease state and smaller patient populations for enrolling in clinical trials and due to disease-related challenges. While rare diseases may take many forms, afflicted patients generally share some characteristics such as rare, costly, and commercially less attractive conditions that obstruct the cost-effective acceptance and implementation of personalized healthcare products. Rare disease therapies are more expensive than therapies for non-rare diseases as a result of several challenges in the drug development process related to these diseases, including systematization and data storage, which are insufficiently utilized in order to maximize costs associated with data analysis. Regulations developed on behalf of rare disease patients may impact a billion people worldwide and often lack commercially attractive market potential. Since 2010, an increasing fraction of new drug approvals has been pharmaceutical patents, and sudden price spikes have received significant interest. Thanks to the Orphan Drug Act, which was signed into law in 1983, this trend has been most favorable for rare disease treatments.

3. Computational Drug Design and Machine Learning

The ability to target disease pathways with small molecules has revolutionized many areas of medicine. However, developing new small molecules requires time, effort, and cost. In this review, we will discuss the process of drug design, and in particular, the application of artificial intelligence to the problem of structure-based drug design. In such a new field, we will see that its current practical application is primarily within the improvement of predictive tools, while state-of-the-art applications are mostly oriented to the assessment of such predictions and knowledge generation.

Computational methods, a well-defined field, are used to reduce time and financial costs of designing new compounds. In their essence, these methods perform the same type of computation as an in vitro biological assay or in vivo animal model, producing a relevant answer. What such methods have in common is either replacing the workforce or increasing the throughput of methods. While clearly the more predictable, reliable, high-throughput computational methods can contribute significantly to developing next-generation pharmaceuticals, they are today still an intricate mix of ambiguity and imperfection. We review here some of the existing tools in compendium as well as the study of the use of machine learning.

3.1. Overview of Computational Drug Design

Computational drug design refers to the use of computer modeling techniques and simulation to discover new drug candidates. It has played an increasingly important role in the drug development process and has witnessed a consistent proportional increase in the number of published research articles. The large number of publications is further supported by robust global investment. The compound annual growth rate of the computational drug design technology services market reached 19.6% between 2017 and 2027. Promising advances in structural biology, bioinformatics, cheminformatics, artificial intelligence, and other fields have driven the sharp increase in computational drug design and created an opportunity to use computational drug design to solve practical problems in rare diseases drug development. In fact, it has been used to complete all stages in the drug development process, including target identification, ADMET prediction, drug repurposing, de novo drug design, generating lead compounds, structure-based and ligand-based drug design, and solving the problems of small scale and complicated pathogenesis in rare diseases drug discovery. A structure-based drug design can identify the promising allopurinol drug candidate for Snyder-Robinson syndrome drug development based on its crystal structure.

3.2. Role of Machine Learning in Drug Discovery

The role of AI in drug discovery has been extensively reviewed. Various AI algorithms have been explored for applications in several stages of drug discovery, such as deep learning as a tool for virtual screening, molecular representation, drug-target interaction prediction, bioactivity prediction, molecule generation, and optimization, as well as de

novo molecule design. Deep learning methods have proven to be highly efficient, accurate, and fast for various drug discovery applications.

AI drug discovery is emerging as a new approach for enhancing existing drug discovery pipelines. With rapid advances and successes, more researchers are focusing on AI-enhanced drug design innovations. Although AI tools and deep learning-based approaches are traditionally developed for easily accessible, publicly available large data sets for diseases with high impact, AI applications in rare disease drug discovery are often overlooked due to the lack of available data. However, the use of AI-driven algorithms in rare disease drug development still has potential. The potential utility of AI in the drug discovery process for rare diseases is mainly embodied in three aspects: 1) data reuse and robustness, where AI can repurpose both clinical and molecular data to guide the mechanisms of action to overcome data scarcity; 2) reducing preclinical times to accelerate the progress of rare drug discovery; and 3) validation before drug entry, which can predict unsuitable drug candidates and reduce the failure rate and time of clinical trials. Although the development of AI applications for rare disease drug discovery still lags behind that of traditional drug discovery, the potential is there, and we anticipate more advanced solutions in the near future.

4. Applications of AI in Rare Disease Drug Development

Initial examples highlight the application of AI in drug development for rare diseases. Rare diseases, with their combined relative frequency of about 6 to 8 per 1,000 live births, remain a challenge for drug development. Until the end of 2014, there were 214 orphan drugs with European marketing authorization, which addressed a total of 109 different diseases, out of approximately 5,000 known ones. Thus, patients with rare diseases remain dramatically underserved. With the current development timelines and high costs for the discovery of drug candidates, the repurposing of drugs, such as one out of four orphan medicines approved that had a new indication or the targeting of these diseases through the development of gene therapy, remains the only viable solutions.

It should be mentioned that the drugs developed through AI-enhanced approaches benefited and may benefit in the future from the incentives provided to companies investing in the development of orphan drugs: Reducing the time of market authorization, Regulatory support for product development, Fee reductions during

development, Protocol assistance, and Reduced risk of not obtaining return on the investment.

4.1. Drug Repurposing

Due to the enormous costs and risks associated with traditional drug development, many researchers turn to the drug repurposing strategy. It is a hypothesis-driven application, but when such a process is performed electronically using relevant bioinformatics tools and molecular databases, it becomes a promising niche for a variety of research on the identification of new therapeutic uses for commercially available drug compounds. Although there are a number of intensity levels, some molecular mechanism types involved in the actual progression of disorders that allow for drug repurposing are common among rare diseases, such as cholesterol transport, cell signaling and transcription, cell rescue, cell maintenance and defense, energy metabolism, and substrate transport. Provocatively, drug repurposing often uses already approved drugs, thus reducing both risk and financial costs compared to the creation of a new drug target.

Animal models are often used to demonstrate the efficacy of repurposed drugs, although this drug discovery platform contributes to the challenging phase between preclinical and clinical translation. Since the use of repurposed drugs is already approved for other indications, the likelihood of them being granted marketing approval by the regulatory agency is increased. This is what happened with the repositioning of a drug for the treatment of pulmonary arterial hypertension. Marketed for erectile dysfunction, the action of this drug for treating this cardiovascular symptom is due to the drug's mechanism of action—it can lead to pulmonary vasodilation by blocking an enzyme playing a role in the chemical cascade that mediates contraction of pulmonary vascular smooth muscle. With more than 300 similar rare disease cases, the approval was granted remarkably quickly, mainly on the basis of animal studies.

4.2. Virtual Screening and Molecular Docking

In the drug design process, it is also common to exploit the well-known three-dimensional (3D) structure of the target protein and then subject it to molecular docking experiments. This approach usually couples the target's 3D structural features with binding affinity assessments of various compounds, usually a set of known or predicted chemical molecules. Notably, this flexible approach allows for the inclusion of modified

small molecules as input samples, which can be performed using mutational fingerprints. Typically, the selection of molecules and molecular docking experiments aims to achieve both improved binding affinity and the utilization of potential pharmacokinetic properties, such as solubility, permeability, variety of metabolism possibilities, and low toxicity.

Developing small molecules as potential ligands for a particular protein is a complex task. Regarding molecular docking, the molecular docking suite is a commonly used and well-tested semi-autonomous tool for virtual high-throughput small molecule screening, which leverages the high-complexity binding principle models of interaction, including both energy-minimized conformers and molecule-protein clashing conformers, for a molecular compound set. Specifically, it generally identifies the fitting 3D position and orientation of small molecules within the binding sites of the chosen target, establishing a high-redundancy cluster analysis and consequently identifying the high-affinity molecules for the target of interest. Due to the steps of receptor and compound preparation, scoring function, and target-specific setups, the user can carry out customized settings to further improve the accuracy of the results. Notably, the scheme also supports new features from newly developed structures in cases where protein conformational states are previously unknown.

5. Case Studies and Success Stories

In 1972, a devastating disease was recognized for a new constellation of symptoms: the body is covered in thick scales, the teeth are misshapen or absent, the hair is sparse, the nails are twisted, and the patient has progressive problems with balance and coordination. In almost all cases, the condition is caused by an alteration of the X chromosome, overactivity of an enzyme called an ADAM protease, and destruction of an important signaling protein encoded at a location on the same gene. Despite a tormented life, painful existence, emotional hardship, and massive financial costs to the patients, there is no cure. For over 40 years, researchers from academia and pharmaceutical companies have attempted to treat this disease. Every reasonable therapeutic approach has been tried, including the development of test compounds designed rationally and using a high-throughput screening strategy. Not one realistic drug has yet been reported. Because the disease touches so few people, it is not a commercial drug discovery target.

Therefore, we used AI to suggest completely new classes of compounds, based on modifying the local charge of the mutated active site, rather than binding inside a very deep natural hydrophobic pocket. We were able to synthesize a panel of these promising molecules, to test multiple ADAM enzymes in the same structure, and to determine the potency of active inhibitors in less than 3 months. The most potent of the inhibitors has an unusual cage structure, with the same thiopurine scaffold as the corresponding top-ranked compound in the reactive center area. The strongest compound is an acid that manifests its inhibitory actions through the alkylation of the two catalyst residues. We herein describe the background of the hereditary disease and this compound, which is a historically significant chemical step.

5.1. Examples of AI-Enhanced Drug Design in Rare Disease Therapeutics

AI and behind AI methods, including computational molecular dynamics, quantum mechanical approaches, molecular machine learning, de novo generation of ligands, and so forth, offer hope in the search for new drug compounds. To accelerate AI-driven advances in rare disease therapeutics, more extensive datasets and more comprehensive data inclusion requirements are needed to ensure equitable representation of all population subgroups in rare disease research studies. Efforts to create more easily accessible and interpretable tools that work out of the box for researchers outside the field of AI for rare diseases will pay dividends in fostering the use of these new tools more widely. The criteria for inclusion in the call to share AI-generated datasets, as well as for journal reviewers, were put in place as a discovery-based approach to open up access to a wide variety of data. Deeper collaborations should be prioritized when creating these AI tools to build competence and capacity within the rare diseases field. While AI may not fundamentally change the basics of drug discovery, the new design methods that AI helps enable may help provide the elusive first leads that have hindered drug discovery in rare diseases. The ease of application and combined suitability of AI and conventional drug discovery resources can pave the way to the more efficient, open, and collaborative processes required for rapidly advancing the field of rare disease therapeutics.

6. Future Direction

Regarding the limitation in the device of AI-enhanced computational drug design, hardware improvement can enhance the capability of the computational process of in

silico modeling. The development of drug carriers and drug delivery technology should be considered for accidents during the process, in imaging and image-guided drug delivery. Besides that, human 3D tissue-organ models and organ-on-a-chip models should be coupled with the computational design process in drug development. The chemical gene analysis in the omics data after exposure to drugs should be analyzed for the accuracy of the mechanism of action prediction, which focuses on the action of drug prediction for computational drug design. A decrease in the small rare disease-related adverse signals in the drug discovery process can actually accelerate drug discovery and support the hype hypothesis or accelerator hypothesis for the study of drug repurposing for rare diseases with more analysis. The optimization step in drug design should be taken for rare disease drugs in order to reduce drug-to-drug interactions for repurposed drugs.

The machine learning technique in the drug-to-drug reaction model should be developed to understand the mechanism of common or high-frequency side effects of the drugs to make decisions in changing the design step for an already identified rare disease drug. Types of rare diseases should be considered in developing the computational drug design model, as the capability of the model to predict one of the three rare disease categories can be different, and the design for rare diseases can be more complex than for common diseases. The collections and designs of rare drug metadata, which are structured labeled rare and non-rare drugs, adverse events for the reaction severity categorization, and the models of drug-gene and drug-reaction pairing can be useful for enhancement. The model predictions, knowledge databases, algorithms, and results should be publicly shared. Besides that, global collaborations in rare diseases and large proprietary companies should be established for the success of the next-generation therapeutics of rare diseases.

7. Conclusion

In conclusion, we propose two major implications of this AI- and computing-enhanced computational drug design roadmap. Firstly, novel and unexpected treatments that have often been overlooked in the scientific study of rare diseases are now possible, whereby second and higher alternatives of functions from biologically related targets are newly revisited as high-confidence inhibitor structures that were ignored or discarded in early- to late-stage projects owing to the limited knowledge of datasets, suboptimal

methodologies, or impact due to a deficiency of sufficient computational data within practical timescales. Secondly, various challenges and fundamental issues related to drug discovery could be uncovered. Our method of evaluating PK properties at the early stage of ligand design will provide dependable human absorption, distribution, metabolism, excretion, and toxicity evaluations that are significant means for accelerating projects. In contrast to target- and data-driven approaches, in this study, molecules with an unprecedented degree of structural diversity and novelty could be discovered with AI models with low data or weak data performance. Our large-scale Structural Activity Relationship analyses serve another ground truth for target- and data-driven approaches to evaluate and benchmark model quality, robustness, and reliability. Moreover, such regulations or rules could also be utilized in the process of developing new optimization models or guiding researchers directly to design new drug-like molecules with better properties based on structural features. Taken together, the proposed method is preparatory and broad-based in the drug discovery stages, helping researchers rationally address each of these challenges.