Efficient graph representation framework for chemical molecule similarity tasks
(Technical project)
Understanding the drug development industry's shift in focus to orphan drugs and its implications
(STS project)
A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Jiaji Ma

[October 27, 2023]

Technical Team Members:
Jiaji Ma

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

MC Forelle, Department of Engineering and Society

Rosanne Vrugtman, Department of Computer Science

Introduction

The field of drug development plays a pivotal role in addressing global health challenges but has faced significant challenges in recent years. Traditional approaches primarily rely on empirical trial-and-error methods, and thus it is typically a decade-long process that is laborious, expensive, and prone to failures, with an overall success rate estimated to be as low as 6.2% (Wong et al., 2018, p. 17). Factors such as increasing regulatory obstacles and the difficulty of finding blockbuster drugs have led to a weakening of research and development productivity in the pharmaceutical industry (Lavecchia, 2019, p. 2017).

Recent advances in technology and growing automation have led to huge amounts of available data in biomedicine, and this has opened new avenues for improving and expediting the drug development process (Paul et al., 2021). It is extremely challenging to effectively utilize these data using traditional approaches, leading to the increasing popularity of machine learning technology in the drug development field. In recent years, machine learning technologies have been moving from theoretical studies to real-world applications and have demonstrated promising results in various fields. There has been increasing application of machine learning technologies in the drug development field, across various stages of the process (Vamathevan et al., 2019, p. 464), since such technologies offer the potential to expedite the drug discovery process by predicting drug properties, gene responses, and more based on molecular structures.

With the aid of vast and ever-expanding data, machine learning has the potential to be a game-changer, with the ability to address some of the most pressing challenges in the drug development field, including accelerating the drug development pipeline, improving product quality, and enhancing cost-effectiveness (Dara et al., 2021, p. 1950). This has the potential to propel the industry forward and provide higher quality healthcare at a lower cost. However,

addressing this problem extends beyond technical innovations. Due to the finite resources that can be dedicated to drug development, equity in healthcare is an ever-growing concern. Thus, it is important to assess recent trends in the industry with regards to the shifting development focus. Understanding the relevant factors leading to this trend and its social impact will help address shortcomings in various aspects.

In the technical topic section, I will discuss my research internship project on computational chemistry at Oak Ridge National Laboratory, focusing on the power of graph machine learning and its application on chemical molecule graphs. By utilizing graph neural networks, we were able to develop models to obtain efficient vector representations of chemical molecules, which can be utilized for downstream tasks. We also demonstrated the superiority of this method compared to traditional cheminformatics tools. In my STS topic section, I will discuss the prospects and challenges in the drug discovery and development process, its ethical and social impacts, especially in the context of machine learning technology utilizations in the field.

Technical Topic

During the summer before my final year of undergraduate studies at the University of Virginia, I had the opportunity to intern at Oak Ridge National Laboratory, as a research intern. Oak Ridge National Laboratory is a federally funded laboratory sponsored by the Department of Energy based in Oak Ridge, Tennessee, which advances research in a variety of scientific fields. The research group I interned under, the Discrete Algorithms group, focuses on areas spanning from graph algorithms in High Performance Computing systems to Machine Learning theory and applications. I worked on a research project proposed by my mentor, who is a research scientist

at the laboratory. The project focused on machine learning and its application in computational chemistry, working with chemical molecules.

In the field of machine learning, the emergence of graph data as a versatile and expressive format has paved the way for tackling complex problems in diverse scientific domains. Graphs have proven to be highly effective for representing and analyzing real-world data in numerous domains, due to their ability to encode both structural and semantic information. By representing entities as nodes and relationships as edges, graphs provide a powerful framework for capturing the underlying patterns, dependencies, and inter-dependencies within complex systems. This flexibility and ability to capture and model arbitrary relationships between arbitrary entities allow graphs to go beyond the limitations of traditional data structures, providing a more comprehensive and holistic understanding of complex systems and real-world data. By modeling chemical molecules as graphs, my technical project aims to leverage machine learning to address challenges in the realm of computational chemistry.

Machine learning has demonstrated tremendous success in a wide array of domains, on 1-dimensional sequential and 2-dimensional grid data, but most algorithms and tasks require the data to be in tensor format, which are sets of algebraic objects related to a vector space. Thus, it is challenging to directly perform machine learning on graphs. Graph representation learning, which converts the raw graph data into vectors while preserving intrinsic graph properties (Chen et al., 2020, p. 2), allows for the efficient utilization of machine learning tools to perform downstream tasks. Various techniques have been developed, Graph Neural Networks (GNNs) in particular have demonstrated promising state-of-the-art performance. There is a wide array of different GNN structures, but they all follow the same general strategy of neighbor aggregation, where each node's feature vector representation is iteratively updated by aggregating

representations of its neighboring nodes (Xu et al., 2018, p. 4), this allows for both structural and functional information to be effectively captured.

In the field of computational chemistry, due to the nature of chemical molecule data, there have been cheminformatics tools for mapping the chemical space long before machine learning, called molecular fingerprinting. Such algorithms iteratively encode circular substructures of a molecule as identifiers, hash them, and fold them to bit positions to generate a bit string (Cereto-Massagué et al., 2015, p. 59), which is a vector that represents the structural information of a molecule. However, they are not able to offer ideal performance due to the length of the resulting embedding vectors.

The research questions my technical project revolves around are the efficient representation of molecular graphs and their applications in computational chemistry. Specifically, how can we transform chemical graphs into computationally efficient and informative tensor format representations that improve the accuracy and efficiency of tasks such as similarity search and prediction of molecular properties?

The technical intervention that will address current challenges and advance the field of chemical molecule representation will be a framework for generating efficient chemical molecule vector representation. By building upon existing technologies by tweaking and combining state-of-the-art models and algorithms, this framework will utilize the power of Graph Neural Networks to accurately predict molecule properties while capturing structural information. The framework will be a powerful tool for the exploration of the chemical space and aid in reducing the time and resources required for discovery of new compounds with specific desired properties. Our findings will provide valuable insights into the application of

machine learning to graph data for chemical molecule analysis and has the potential to bring more effective life-saving drugs to market in a timelier manner.

STS Topic

The drug development industry has produced medications that have greatly increased quality of life for many, both through ameliorating pain and treating diseases. The 1990s was marked by an innovation boon in drug developments and various blockbuster drugs' success, which led to skyrocketing market growth, primarily fueled by interest in drugs targeting chronic diseases with potential for large financial returns (Osakwe & Rizvi, 2016, p. 8). This marked the beginning of incremental shifts in the industry. Historically, patients with rare diseases have been underserved by drug development. Recently there has been legislation in various countries to incentivize and encourage the development of drugs to treat rare diseases (called orphan drugs), to address this disparity (Haffner, 2006, p. 446).

Equity, especially access to treatment, is an important principle in healthcare. Patients suffering from rare conditions should be entitled to the same opportunity of receiving treatment as anyone else. However, there are currently still serious shortcomings in the development process, as standard procedures have suboptimal performance in capturing societal values when evaluating drugs, as orphan drugs do not prove to be cost effective in most cases. (Drummond et al., 2007, p. 36). Additionally, due to the small market, there are still many obstacles for individual patients in this regard, notably limited access.

Thus, more research is required to gain a better understanding of the trends in the drug development industry, which will allow for the shortcomings to be addressed better. A key research question I aim to answer using this STS project is how have various sociotechnical

factors contributed to this shift of focus towards orphan drugs in the drug development industry. Answering this question will mainly rely on reviewing existing literature. I will review studies in medical journals that introduce the pharmaceutical ecosystem, drug development process, the stakeholders, and how it interacts with the overall society. Doing this will help provide insights on the evolving landscape of the drug development industry and how it impacts society. Additionally, from the aspect of government intervention and societal value, I will examine works in sociology studying the impact of government initiatives and discuss problems with technology assessment and the bias in societal value evaluation.

The research on how the drug development industry has transitioned towards focusing on orphan drugs will be conducted in the framework of technological determinism, which is the belief that technology is an important governing force in society and drives social change, while some critics argue this fails to account for the complex relationship between technology and society (Smith & Marx, 1994, p. 2). This framework is appropriate since the question of interest involves the impact of technological advancements in the drug development field on social change, especially in the context of the underserved population of patients with rare diseases.

Conclusion

In my technical project, I successfully developed a model that can produce efficient vector representations of chemical molecule graphs, so that they can be efficiently utilized in downstream tasks, specifically aid in the identification of potential drug molecules. In my STS project, I will be investigating how sociotechnical factors have driven the pharmaceutical industry's transition to focusing on rare drugs, and the social impact. This can lead to informing policymakers of more effective regulation, guiding pharmaceutical companies in strategic adaptation, and empowering patients through increased awareness and advocacy for equitable

access to treatments. The combination of these projects can help address the need for more effective and equitable drug development processes, since the STS project assesses the sociotechnical factors behind the recent trends in the drug development industry and the technical project proposes a tool for improving the drug development process.

References

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods, 71, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005

Chen, F., Wang, Y.-C., Wang, B., & Kuo, C.-C. J. (2020). Graph representation learning: A survey. *APSIPA Transactions on Signal and Information Processing*, *9*(1). https://doi.org/10.1017/atsip.2020.13

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2021). Machine learning in Drug Discovery: A Review. *Artificial Intelligence Review*, 55(3), 1947–1999. https://doi.org/10.1007/s10462-021-10058-4

Drummond, M. F., Wilson, D. A., Kanavos, P., Ubel, P., & Rovira, J. (2007). Assessing the economic challenges posed by Orphan Drugs. *International Journal of Technology Assessment in Health Care*, 23(1), 36–42. https://doi.org/10.1017/s0266462307051550

Haffner, M. E. (2006). Adopting orphan drugs—two dozen years of treating rare diseases. *New England Journal of Medicine*, 354(5), 445-447.

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

Osakwe, O., & A., R. S. A. (2016). *Social aspects of drug discovery, development and commercialization*. Elsevier / Academic Press.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1), 80. https://doi.org/10.1016%2Fj.drudis.2020.10.010

Smith, M. R., & Marx, L. (Eds.). (1994). Does technology drive history?: *The dilemma of technological determinism.* MIT Press.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

Wong, C. H., Siah, K. W., & Lo, A. W. (2018). Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2), 273–286. https://doi.org/10.1093/biostatistics/kxx069

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How Powerful are Graph Neural Networks?. *In International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.1810.00826