Deep Multimodal Representation Learning to Integrate Natural Language with Genomic Interval Data for Tailored Biomedical Discovery
Evaluating Social Pharmaceutical Innovation as a Means of Reducing High Therapeutic Costs

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 Biomedical Engineering

By
Lilian Jones

November 19, 2023

Technical Team Members:
Caitlyn Fay, Peneeta Wojcik, and Zach Mills

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

Dr. Kent Wayland, Department of Engineering and Society
Dr, Nathan Sheffield, Department Public Health Sciences, Biomedical Engineering, Biochemistry and Molecular Genetics, and Data Science
Nathan LeRoy, Graduate Student, Department of Biomedical Engineering

General Research Problem: Addressing High Prices in Pharmaceutical R&D for Rare Disease Therapies

What can be done to reduce the costs of research and development (R&D) and, therefore, the prices of therapies treating rare disease?

The R&D necessary to bring a new pharmaceutical to market is an expensive and long-term investment. From 2009 to 2018, the average cost was estimated to range from $1042.5 million to $1637.5 million, and these expenses continue to rise today (Morgan et al., 2011). Typically, in the R&D process, clinical trials are the most laborious and costly stage (Parker-Lue et al., 2015). In the case of developing therapies for the treatment of rare diseases, the costs of the clinical trials are even greater. One factor contributing to this is, because the disease is rare, there is a much smaller patient population as compared to more common ailments. Thus, more resources must be invested into recruiting participants and gathering data for the trials. The complications of heightened R&D costs and smaller market sizes for these rare diseases lead pharmaceutical companies to charge high prices to patients and insurance companies in order to make a return on their investment. By addressing the lack of patient data available to researchers, this has the potential to help bring down the cost of R&D and incentivize pharmaceutical companies to then reduce the prices of their therapies and improve access to these life-saving medications for patients. A way to tackle this deficit of data is to use machine learning methods to train models on existing data to be able to generate novel data and increase the breadth of resources available for investigators to use.

For the technical portion of this research, my capstone team and I will attempt to create a generative model that uses natural language processing (NLP) techniques to input a search query, such as “pediatric cancer” or “glioblastoma”, and outputs relevant genomic interval data that can

then be used in subsequent biomedical experimentation. We will approach this task through trying out four different model designs: text-to-BED neural network, direct encoder, diffusion model, and transformer model, then selecting the best performing model to move forward with in training. In addition to this technical approach of generating data, a more theoretical approach is to use the research concept of social pharmaceutical innovation (SPIN) to disrupt the current, rigid and highly regulated framework of pharmaceutical innovation to alleviate the burden of high therapeutic costs on patients. In my science, technology, and society (STS) research I will explore how promoting collaborations between the public and private stakeholders in the pharmaceutical industry will streamline regulatory processes and promote research for rare disease therapies. Supporting rare disease research will thus help to reduce drug prices for patients by incentivizing pharmaceutical companies to price their products with these patient's best interests in mind, not just for profits.

Technical Research Question: Deep Multimodal Representation Learning to Integrate Natural Language with Genomic Interval Data for Tailored Biomedical Discovery

How Can a Machine Learning Model be Used to Integrate Natural Language with Genomic Interval Data to Generate Novel Data Sets?

The amount of data from epigenomic sequencing experiments has exploded over the past 10 years, increasing exponentially as technologies continue to improve. This large volume of data exists, because, although the human genome has been fully sequenced, even cells with identical genetic material can have vastly different phenotypes. Understanding and analyzing how these variations in phenotype occur and what genomic modifications cause them is an important consideration in the field of epigenomics today. In fact, failing to adjust studies for

cell-type heterogeneity can limit the accuracy and sensitivity of sequencing technologies to locate these modifications (Li et al., 2022). Thus, the development of such an overwhelming amount of sequencing data has created a clear demand for complex models to understand the genomic relationships that exist within this large volume of data (Gharavi et al., 2023). One prospective approach is using the machine learning concept of multimodal representation learning for handling such an expansive volume of data.

Multimodal representation learning is a type of machine learning focused on training neural network models using multiple data sources, and has been applied successfully in genomics research (Gharavi et al., 2023). Experimental data from multiple sources can be combined to extract the most important features of each data set. Choosing to use a multimodal model compared to a unimodal model gives way to more dynamic predictions, and therefore better performance of the model. At present, the models created for genomics research are based on Word2Vec, a context-aware neural network model used to learn word associations (Mikolov et al., 2013). The neural network mathematically transforms natural language text into vectors of numbers so that the similarities and differences between words can be calculated and understood by the computer. We want to use this same machine learning technique of transforming large quantities of data into low dimensional vectors that can be interpreted by an algorithm by working with genomic data in the form of Browser Extensible Data (BED) files (*Genome Browser FAQ*, n.d.). The goal of our technical research is to utilize assay for transposase-accessible chromatin with sequencing (ATAC-seq) data and chromatin immunoprecipitation sequencing (CHiP-seq) data to create four distinct deep-learning models that generate relevant genomic region sets to a user-entered search.

The outputted genomic interval data, formatted as BED files, will contain information such as the location of genes, transcription factors, chromatin accessibility, and methylation that can then be used for in silico bioinformatics studies. The four models: adapted text-to-BED, direct encoder, diffusion, and transformer techniques, will be generated in Python and will be able to represent genomic region data as low-dimensional dense vectors called embeddings, with the most effective and accurate model being selected for future use after validation. Next, we will train a feed-forward neural network to relate the vector embeddings of the metadata text of the BED files to the embeddings of vectors from the genomic data in the BED files. Finally, we will create an interface that takes in text (specifically a query pertaining to the biomedical area of interest), and returns the most relevant genomic region to this search. This is achieved by transforming the text description to an embedding, running the embedding through the comparison model, and decoding the output region embedding into genomic region data. The interface will be hosted on public domain to allow for widespread use of the model by biomedical researchers.

The purpose of creating four models is to approach the design from varying machine learning principles to find the most accurate model. The different models: a text-to-BED file neural network, a direct encoder, a diffusion model, and a transformer, all range greatly in methods, execution, and model complexity. By creating each model and determining which of the four has the best performance through training and testing, we will be able to generate genomic interval data for user-entered queries. We will assess model accuracy through running the k-Nearest Neighbors (kNN) algorithm and calculating the Silhouette scores of the resulting clustering. The model with the highest Silhouette score, indicating the model with the greatest ability to discern the similarities and differences of various BED files, will be selected.

In the end, the development of a successful model will be particularly powerful for the biomedical research of pathologies where experiments were previously hindered because of limited data. Whether the limitations were due to the rarity of the condition or because of inaccessibility of the relevant tissue for sampling, generating BED files from our deep representation learning model will advance bioinformatic and computational genomic research in these areas. It will also reduce the human error that comes from manually extracting and combining multiple sets of genomic data. Furthermore, by creating a user-friendly interface for entering specific queries into the model and releasing it on a public domain will make the interface widely available to biomedical researchers and democratize its use.

Evaluating Social Pharmaceutical Innovation as a Means of Reducing High Therapeutic Costs

How can SPIN be used during the pharmaceutical R&D process to shift the burden of costs away from patients while maintaining innovation incentives for pharmaceutical companies?

By providing public use of the machine learning model, the goal is to allow researchers to use the model to generate data necessary for the R&D pipeline, particularly for rare diseases where accessibility to data is limited. While this scientific approach to addressing the extremely high costs of rare disease therapies is promising, it is still necessary for efforts from pharmaceutical and insurance companies as well as the federal government to reduce prices for patients. Thus, in my STS research I want to examine how the research theory of Social Pharmaceutical Innovation (SPIN) can be used to reframe how the industry looks at the drug development process and lead to lower prices for patients.

Currently, corporations in the pharmaceutical industry operate in such a way that prioritizes making a return on their investments. While this mode of operation may be sufficient

for other industries, the products developed by pharmaceutical companies are non-trivial and are often life-saving therapies. Being motivated to develop new products, especially for those treating rare diseases, based solely on profits ultimately hurts patients as these products are priced highly to offset the high cost of R&D and the smaller market sizes. Additionally, a pharmaceutical company may be motivated at the prospect of being the sole supplier of a rare disease drug in the market and thus invest in their development. However, this leads to a lack of competition, allowing the company to practically set the price at whatever they would like and stripping away any incentive to continue to innovate and increase the quality of the drug. Furthermore, current practices in pharmaceutical innovation divides the industry into distinct sectors. The strict regulatory framework of approving drugs discourages collaboration between fellow pharmaceutical corporations and as well with the public sector. Attempting to combat this and promote collaboration will help to create harmonization between the stages of R&D and drug approval and between the public and private industries by streamlining research and pooling together resources (Siddiqui & Rajkumar, 2012).

SPIN is defined as the use of novel forms of collaborative approaches, initiatives, policies, methods, and/or designs engaging various stakeholders that diverge from traditional pharmaceutical innovation practices (Douglas et al., 2022). Thus, the purpose of SPIN is to drive the creation of safe, efficient, and readily available therapies, catering to unmet needs of rare disease patients, prioritizing social impact over market-driven motives.

As previously mentioned, this concept involves various stakeholders, including these pharmaceutical companies, researchers, healthcare providers, patients, advocacy groups, and governments. The goal is to have these groups working together to address unmet medical needs and improve access to safe, effective, and affordable medications. SPIN is a direct response to

some of the criticisms and challenges associated with the pharmaceutical industry, including high drug prices, lack of access to essential medications, and ethical concerns. By emphasizing collaboration, transparency, and a commitment to improving the well-being of patients and society, social pharmaceutical innovation aims to create a more sustainable and equitable healthcare system. This is a very new concept, introduced in 2022 through combined efforts of researchers from areas of social sciences, law, and public health (Douglas et al., 2022). Since this is still a novel framework, it is important for research such as this to assess its potential in helping to disrupt the current model of pharmaceutical innovation.

In addition to placing importance on open collaboration and public-private partnerships, other SPIN strategies include shifting to a model of awarding prize funds instead of grants to reward pharmaceutical companies who successfully bring drugs to market, removing the incentive to price the therapies at high rates. Likewise, creating better harmonization between companies and various regulatory organizations will streamline the rigid regulatory processes and encourage investment in drug development (Rollet et al., 2013). Also, using compulsory licensing to grant pharmaceutical companies to develop generic versions of these expensive therapies. This will increase competition in the market and thus reduce prices. Negotiating agreements between the government and pharmaceutical companies to set drug prices for a certain period after market entry will allow companies to make a return on their R&D expenses while still ensuring affordability. The government can further provide tax incentives as well to pharmaceutical companies, making it more financially viable for companies to invest in the development of treatment for rare diseases. Moreover, promoting collaboration beyond just US organizations, but to a global scope can be useful in advancing research efforts. By merging resources, sharing data, and again harmonizing regulatory processes, the development process

will accelerate, thus reducing costs However, none of these strategies will be entirely effective without heavily involving patients and advocacy groups in the drug development process. By emphasizing their insights, priorities can be set for research, ensuring that their pressing needs are valued and addressed (Dranove et al., 2014).

Ultimately, SPIN can be used to frame the operations of many relevant stakeholders in the drug development process. Regulatory bodies such as the Food & Drug Administration (FDA) and the European Medicines Agency (EMA) can collaborate to model their processes after one another to streamline testing and approval of new products, thus helping to decrease R&D prices for pharmaceutical companies. Likewise, increasing communications and transparency between these regulatory agencies and pharmaceutical corporations will further help in streamlining the approval process and reduce developmental expenses. Furthermore, pharmaceutical companies can apply SPIN to their company ethos to promote collaboration with competitors to harness resources and aid each other's research efforts to advance the development of life-saving therapies and reduce costs for patients.

The resulting impact of implementing SPIN methods can be analyzed through indices such as health outcome metrics, accessibility and affordability metrics, economic impact metrics, and R&D innovation metrics. Since SPIN is still a novel concept, it is not yet enforced as a standard of operation for pharmaceutical companies by regulatory bodies. Thus, computing these metrics will give insight on the performance of the companies who are adhering to SPIN. For quantitatively assessing how SPIN can impact patient outcomes, patient surveys can be conducted to identify how mortality rates, life expectancy, quality of life, and patient reported outcomes (PROs) have changed throughout the time since adopting SPIN principles. Similarly, these patient surveys can also give insight into the economic impact of a therapy developed

under SPIN principles by calculating the cost per unit of health improvement compared to alternative therapies as well as calculating how much patients have saved in healthcare costs due to the innovation. Geographically assessing the distribution of the therapy will provide a useful metric for understanding how accessible the product has become as compared to previous therapies. Additionally, the affordability of the innovation can be determined by calculating an affordability index, the price of the therapy in relation to income levels and healthcare expenses. In terms of assessing success from an industry standpoint, the change in the amount invested in R&D by both a pharmaceutical company and government agencies will illustrate the impact on innovation. Likewise, quantifying the number of patents and new molecules a pharmaceutical company has produced since adopting SPIN will demonstrate the improvement in the robustness and diversity of their innovation pipeline (Fellows & Hollis, 2013). In summary, these metrics can be used to holistically understand the impact of SPIN in the development of a new therapy and to hold pharmaceutical companies accountable for its effective implementation while efforts to make SPIN a legislative standard are underway.

After first completely outlining the concept of SPIN, in my STS paper I will then further elaborate on the strategies listed above as well as additional approaches as to how theoretically they can be implemented by the stakeholders of the pharmaceutical industry to lower drug prices. This will be achieved by reviewing the current literature on SPIN and comparing the proposed improvements as compared to current practices in the industry. Specifically, to evaluate SPIN in a more empirical mode, I will assess how value-based pricing can be applied to establish new pricing models for these therapies. To elaborate, value-based pricing refers to the idea of pricing drugs based on the actual value they bring in terms of improved health outcomes and quality of life. By computing the value-based price of a therapy for a rare disease and comparing this to the

current industry prices, it will give valuable insight into how SPIN can be used to disrupt conventional pricing models and to meet patient needs at more affordable rates. I will use fragile X syndrome, cystic fibrosis, hemophilia, and juvenile idiopathic arthritis as specific empirical case studies, analyzing data from cost-of-illness studies done on these rare diseases (Armeni et al., 2021). These are all examples of rare diseases that have extensive economic and epidemiologic data and thus have been assessed in various cost-of-illness studies. Understanding how the cost of these rare diseases beyond just monetary terms can be calculated will provide a useful framework of how cost-of-illness studies can be performed and integrated into the research considerations of drug development.

Conclusion

Through looking at specific case studies of diseases where the concepts of SPIN, particularly value-based pricing, can be applied, I want to assess how this new research technique can be used to approach the issue of high therapeutic costs. The drug development pipeline is an expensive, rigid, and lengthy process in the US. Lack of harmonization between the various regulatory agencies and the companies producing these drugs is a complicated roadblock in the R&D process (Parker-Lue et al., 2015). Thus, the prospect of using principles of SPIN to improve transparency and collaboration is a promising avenue toward resolving this issue. On the other hand, a more technical approach to solving this issue is addressing the large expense that comes with gathering patient data for clinical trials. Recent advancements in machine learning and artificial intelligence can be leveraged to develop models to address just that. A complex neural network can then be trained on existing data in order to generate novel data to be used in silico bioinformatic analysis. Additionally, integrating natural language

processing concepts will help to create a model that is accessible and easy to use (Gharavi et al., 2023). By creating a generative tool, less resources will need to be invested in recruiting participants and gathering data for clinical trials, thus helping to reduce the expenses of drug R&D. In summary, taking both a technical and social approach to creating incentives for pharmaceutical companies to reduce costs of therapies meant to treat rare diseases is a robust means of ensuring the delivery of these life-saving drugs to patients while still promoting innovation in the industry.

References

Armeni, P., Cavazza, M., Xoxi, E., Taruscio, D., & Kodra, Y. (2021). Reflections on the Importance of Cost of Illness Analysis in Rare Diseases: A Proposal. *International Journal of Environmental Research and Public Health*, *18*(3), 1101. https://doi.org/10.3390/ijerph18031101

Douglas, C. M. W., Aith, F., Boon, W., de Neiva Borba, M., Doganova, L., Grunebaum, S., Hagendijk, R., Lynd, L., Mallard, A., Mohamed, F. A., Moors, E., Oliveira, C. C., Paterson, F., Scanga, V., Soares, J., Raberharisoa, V., & Kleinhout-Vliek, T. (2022). Social pharmaceutical innovation and alternative forms of research, development and deployment for drugs for rare diseases. *Orphanet Journal of Rare Diseases*, *17*(1), 344. https://doi.org/10.1186/s13023-022-02476-6

Dranove, D., Garthwaite, C., & Hermosilla, M. (2014). *Pharmaceutical Profits and the Social Value of Innovation* (w20212; p. w20212). National Bureau of Economic Research. https://doi.org/10.3386/w20212

Fellows, G. K., & Hollis, A. (2013). Funding innovation for treatment for rare diseases: Adopting a cost-based yardstick approach. *Orphanet Journal of Rare Diseases*, *8*(1), 180. https://doi.org/10.1186/1750-1172-8-180

Genome Browser FAQ. (n.d.). Retrieved November 19, 2023, from https://genome.ucsc.edu/FAQ/FAQformat.html#format1

Gharavi, E., LeRoy, N. J., Zheng, G., Zhang, A., Brown, D. E., & Sheffield, N. C. (2023). *Joint representation learning for retrieval and annotation of genomic interval sets* (p. 2023.08.21.554131). bioRxiv. https://doi.org/10.1101/2023.08.21.554131

Li, Y., Ma, M. X., & Renneboog, L. (2022). Pricing art and the art of pricing: On returns and risk

in art auction markets. *European Financial Management*, *28*(5), 1139–1198. https://doi.org/10.1111/eufm.12348

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., & Greyson, D. (2011). The cost of drug development: A systematic review. *Health Policy (Amsterdam, Netherlands)*, *100*(1), 4–17. https://doi.org/10.1016/j.healthpol.2010.12.002

Parker-Lue, S., Santoro, M., & Koski, G. (2015). The Ethics and Economics of Pharmaceutical Pricing. *Annual Review of Pharmacology and Toxicology*, *55*(1), 191–206. https://doi.org/10.1146/annurev-pharmtox-010814-124649

Rollet, P., Lemoine, A., & Dunoyer, M. (2013). Sustainable rare diseases business and drug access: No time for misconceptions. *Orphanet Journal of Rare Diseases*, *8*, 109. https://doi.org/10.1186/1750-1172-8-109

Siddiqui, M., & Rajkumar, S. V. (2012). The High Cost of Cancer Drugs and What We Can Do About It. *Mayo Clinic Proceedings*, *87*(10), 935–943. https://doi.org/10.1016/j.mayocp.2012.07.007