

Thesis Project Portfolio

Understanding AlphaFold and Its Implications for a Deep Learning Approach to Protein-Compound Modeling

(Technical Report)

The Ethical Ramifications of the Commodification of DNA

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Meghan Flannery Anderson

Spring, 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Understanding AlphaFold and Its Implications for a Deep Learning Approach to Protein-Compound Modeling

The Ethical Ramifications of the Commodification of DNA

Prospectus

Sociotechnical Synthesis

Advancements in bioinformatics in recent years have opened the door to solving long-standing challenges in the development of pharmaceutical drugs. Namely, machine learning advancements have revolutionized the ability for researchers to predict how a certain drug will interact with proteins. These computational models rely heavily on expansive databases to deduce protein structures, which require genetic information. The genomic databases in play are of particular interest as the sources of genetic data and the means of obtaining rights to that data pose privacy issues, among other concerns. In order for research in the field to continue equitably, a reassessment of the collection of genetic data is necessary.

Current trends suggest that many institutions are leaving traditional experimental approaches behind in favor of less expensive, less time-intensive computational efforts. In 2020, the Alpha Fold program was released and hailed as a major scientific breakthrough for predicting 3D conformations of proteins. Generating highly accurate predictive models for how certain compounds will interact with proteins may soon be possible. The technical thesis of this body of work investigates both the Alpha Fold program and contemporary methodology for predictions of compound-protein interactions (CPI). The literature review serves as a starting point towards the end of creating an effective predictive model for 3D visualization of CPIs based on the successes of Alpha Fold.

One source of information for genetic databases handled in pharmaceutical research and development lends itself to the sale of data by direct-to-consumer genetic testing companies to pharmaceutical companies. Three major ethical concerns arise from these alliances: (1) whether informed consent for research conducted on DNA is obtained by agreements signed in the initial collection of data, (2) whether consumers have sufficient knowledge of the implications of the storage and sale of their DNA, and (3) whether therapies derived from the DNA of an individual are eligible for patenting. The STS thesis situates the rights of individuals in the modern market for DNA.

It should be noted, the technical prospectus was developed in Fall 2021 and related to a project for predicting gene expression based on histone modifications. When progress stalled on the project at the start of the Spring 2022 semester, my project pivoted to independent research into protein modeling. Initially, I had high hopes for making some modifications to the Alpha Fold program to enable some semblance of compound-protein binding prediction. Further research into the area proved that such an assignment was unlikely to be completed in the allotted timeframe with the given resources. A literature review was produced instead.

Nevertheless, I've accumulated significant knowledge on machine learning practices in the research area. I look forward to continuing to learn more about both research areas and intend to keep an eye out for future legislative and regulatory measures in genomics. Finally, I wish to express my gratitude for my advisors, Dr. Sean Ferguson and Dr. Yanjun Qi, without whom these works wouldn't be possible.