I am pursuing doctoral study because I want to make meaningful contributions to biomedical science in the way I'm best capable – through applied mathematics and computer science. Early on in my undergraduate career, I joined a research team working on developing a neural patch for traumatic brain injuries. During the course of this research, I came to appreciate the enormous gaps in our knowledge of key biological processes. I am fascinated by this uncharted territory. However, I always had a lingering feeling that I was never doing enough; the scale of my work seemed too small. I was drawn to the more immediate and far-reaching prospects of computer science and mathematics. My undergraduate classes, internships in industry, and current job as a software developer at [redacted] have all given me invaluable training in developing algorithms, building data pipelines, and employing ML and AI techniques to gain insights from user data. Now, I'd like to redirect my efforts towards the many new and impactful areas of biomedical science that are best explored using computational techniques.

The first time I dealt with big data was during a summer 2018 internship at [redacted], working with the company's [redacted] search engine. I created a pipeline that processed the entirety of [redacted]'s search database and calculated relevant statistics about the queries. I developed a relational database and pipeline that processed the large datasets and was scalable. I modified the data types and reduced accuracy, where appropriate, to improve storage costs. I had heard the term 'big data' in the past, but I didn't quite conceptualise it until I was tasked with creating a pipeline that processes every search query input into [redacted] yet requires minimal maintenance. I spent most of my summer experimenting with different techniques that saved on processing time and data storage.

The next summer, I took part in a second internship where I was tasked with developing an ML classifier for [redacted] that could determine whether a set of images was related to a given search query. When I was first assigned the task, I wrote it off as simple – I thought I could just feed

information about the images or the images themselves into a few black box ML models and see which model produced the best results. I quickly realised, though, that it was not that easy. Because my classifier needed to run in less than a few milliseconds, I was unable to use AI techniques on the images. I dove into the field of Natural Language Processing. I incorporated features that mimicked how humans would compare items rather than expecting the computer to learn how to do so on its own. For instance, I calculated scores that ranked how related the keywords in a search query were to the keywords in image URLs. I developed hundreds of such features and incorporated them into a gradient boost classifier that ran quickly enough to satisfy the run time constraints. I was still looking for ways to optimise my classifier, so I read the literature and tested out some modifications to gradient boost models. I ended up employing Multiple Additive Regression Trees, as they seemed to improve my overall performance.

Through this experience, I learned that there was so much more to using computational techniques than knowing how to run ML tools. I needed to test out innovative solutions and employ techniques from unexpected subfields of computer science and mathematics to achieve better results.

At work, I am currently leading a project involving the modification of an algorithm that estimates statistics relevant to a specific type of exercise. I am tasked with both reducing the algorithm's computational cost and generating accurate statistics with drastically less input data coming from motion sensors. I conduct tracing analyses and CPU sampling to analyse the performance of algorithms. I then apply mathematical and computational techniques to reduce power consumption. For example, I use interpolation to account for sparse data and modify mathematical calculations to reduce occurrences of matrix multiplication. I truly enjoy the trial and error of exploring creative solutions to data challenges, and I want to leverage my knowledge of data pipelines and processing to the field of biomedicine.

I am most interested in *Theme 1: Computational & Data-Driven Structural Approaches in Drug Discovery*. I believe my background in

mathematics, software engineering, and data processing will allow me to work effectively towards rational, data-driven drug design. Dr. Deane's group's use of statistical and computational approaches to the study of protein structures and drug development aligns with my interest in computational techniques. I am very intrigued by Dr. Deane's approach to simplifying the mathematical problem of protein structure discovery by looking at how the protein folds as it is produced. Dr. Holmes's research in statistical ML also caught my eye since there is so much room for exploration and creativity when developing ML models for data such as MRI images, clinical data, and genomic data. Finally, I am interested in Dr. Minary's research in algorithm and software development in computational structural biology.

This program would help me apply the knowledge I've gained in industry to something I truly care about—harnessing data to produce actionable insights into human health. I know that I would benefit greatly from a strong educational foundation in biomedicine afforded by Modules such as *Molecules, Cells, and Systems* and *Mathematical Model Building in Biomedicine*. The initial taught phase and the two exploratory research projects would give me the necessary context with which to decide where I should direct my efforts. I hope to use my industry training in software development practices in collaboration with the program's industry partners. The program's open-source Software Projects would additionally give me the important experience of working collaboratively to produce accessible biomedical tools. I believe your program would provide the exemplary education, robust network, cutting-edge computing experience, and opportunity for research that I need to build a successful and meaningful career in the biomedical sciences.

Word count: 1000