Witnessing my family member struggle with diabetes and cancer firsthand helped me understand how illness can limit an individual's physical and mental freedom. Also, it raised a question within me. "Will there ever come a day when humans live without suffering from disease and can fully enjoy their lives?" This big question has guided my journey—from studying bioengineering during my master's to working as a machine learning researcher in healthcare startups, where I developed predictive models to enhance medical practice. However, I came to realize that while we can predict many diseases, definitive treatments remain elusive. This realization fueled my interest in bioinformatics, a field that explores disease mechanisms and advances therapeutics through data. I hope my graduate study will guide and deepen my research in bioinformatics, serving as a milestone in my research journey.

Over the past seven years, I have worked as a machine learning researcher at two healthcare startups, where I developed disease screening products using biosignals and evaluated their effectiveness in clinical practice. I enjoyed every moment of my career, especially since it was satisfying to see how our efforts resulted in products that provide help to people. Throughout the process of product development, I encountered many technical challenges, including domain generalization and noisy labels. Domain generalization, in particular, was crucial to ensure reliable performance in the real world. To address this, I particularly focused on data augmentation and structural inductive biases.

One approach I explored was data augmentation to simulate demographic distribution shifts in ECG data. The question was, "How to simulate demographic variations in ECG?" To find a hint, I first studied the principles of ECG signals and found that demographic factors cause differences in heart orientation, position, and chest size, which in turn affect ECG. This insight led me to think of augmentation based on these. Since ECG records the heart's electrical activity from multiple leads, each providing different views of the heart, changes in the heart's orientation or position are equivalent to altering the views. Based on this concept, I simulated changes in heart orientation and position by perturbing these views. Specifically, I modeled a graph representing the relationships between ECG leads, perturbed these relationships, and used the perturbed graph to generate synthetic ECG data. This approach achieved a performance improvement of 3% across various datasets compared to existing methods.

While effective on some datasets, this approach raised questions about its ability to address the data distribution shift the model struggles with, as the transformations were applied randomly rather than targeting specific distributions. To overcome this limitation, I explored adversarial data augmentation (ADA), which generates data distributions that are challenging for the current model. However, I found that existing ADA couldn't simulate temporal changes in ECG. Given that temporal characteristics vary across different demographic groups in ECG, I believed addressing this aspect was crucial to fully tackling the distribution shift problem. To solve this, I proposed differentiable time warping, a method that incorporates a time-warping algorithm into ADA but leverages the frequency domain to overcome the non-differentiability of time warping. To be specific, I first demonstrated that when the signal is divided into small segments and each segment is shifted, their collective effect is equivalent to time warping. Here, we can represent each shift as a phase shift in the frequency domain, where phase shifts involve differentiable addition operations. The synthetic data generated by the proposed approach during training effectively addressed out-of-distribution problems. This method complements ADA by causing the perturbation that ADA cannot simulate, providing additional benefits. As a result, combining both methods resulted in a 40% performance improvement in F1 scores on a specific dataset. UMAP analysis confirmed that the proposed method effectively addressed demographic distribution shifts in ways distinct from ADA, validating my original objective.

In addition, I explored the structural inductive biases of neural networks and their impact on improving generalization to unseen data. Unlike in the computer vision domain, one intriguing observation for the ECG domain was that convolution-based models often outperformed pre-trained transformers, regardless of the size of the data. This led me to question, "Can representations from convolution-based models guide transformers more effectively than self-supervised learning?" To find the answer, I trained a transformer using a block-by-block knowledge distillation method, where each block of a convolutional

neural network guided the corresponding transformer block. This approach allowed the self-attention of the transformer to acquire properties like translational invariance and locality. Interestingly, these properties were not evident when knowledge distillation was performed using the logits. Additionally, transformers trained with this approach outperformed convolutional networks.

While it was gratifying to see my disease prediction and diagnostic products used in medical settings, I also realized that even when diseases are diagnosed early, often there are no definitive treatments available. For instance, despite developing a model for the early detection of heart failure using ECG, I found that treatment often focuses merely on slowing the disease's progression, as no cure exists for conditions like left ventricular diastolic dysfunction. Similarly, diseases like diabetes and hypertension are managed long-term rather than cured, which has driven my interest in exploring the fundamental causes of diseases and developing effective treatments, leading me to the field of bioinformatics.

AI has already made a significant impact on bioinformatics, yet technical challenges still persist in this field. For instance, bioinformatics data are collected under varied experimental conditions and methods, which can affect the quality and consistency of the data. Thus, training models robust to variations arising in experiments is crucial. Setting appropriate structural inductive biases is also important, requiring insights that reflect the characteristics of different data types, especially when dealing with complex and high-dimensional data. Label noise is also prevalent, potentially causing models to learn incorrect patterns. Reducing their impact during training is essential. In fact, these challenges mirror those I faced when applying AI to medicine, and I hope to use my experience to address them.

By tackling the technical challenges of applying AI to bioinformatics, I aim to further advance its application in this field. Based on this, I aspire to address a variety of bioinformatics challenges. I am particularly interested in enhancing therapeutic targeting from gene to protein through the lens of AI. In this regard, I hope to explore genomics and transcriptomics, which are essential for identifying key genetic drivers of disease. Additionally, I am interested in investigating structural biology, including protein-protein interactions and protein design, which can offer immediate therapeutic effects.

Looking back on my time at two startups, I often found that the demands of product development left me with limited time to dedicate solely to research. Nevertheless, I made an effort to carve out time for it, which reflects my deep passion for research. Now, I aspire to return to academia, where I can fully dedicate myself to research and study. After graduation, with a deeper understanding of bioinformatics and AI, I aspire to make meaningful contributions toward a world where humanity no longer suffers from disease.

I believe I have the assets needed to succeed in the graduate program. My academic background includes engineering, computer science, and biology—foundational fields that will enable me to quickly absorb and apply new knowledge. While I know there is much to learn in this field, I have a genuine passion for learning; driven by curiosity, I completed 185 credits in my undergraduate studies, well beyond the 130 required for graduation. Moreover, working in startups, which often involve uncertainty and an overwhelming workload, has taught me valuable lessons: the ability to navigate uncertainty without a manual and to think independently. Last but not least, I experienced success. Although it was a mini-success, as the saying goes, "It takes one who has tasted success to believe in success", I will continue to strive for further success.