Will there ever be a day when humans live without suffering from disease and fully enjoy their lives? How can we utilize AI to liberate humanity from illness? Witnessing firsthand a family member suffer from cancer and diabetes, and understanding how illness can limit an individual's physical and mental freedom, has deeply motivated me to pursue these questions. During the PhD studies, I aim to take steps toward finding answers. Specifically, I would like to investigate applying AI to biology and medicine to deepen our understanding of disease and contribute to the development of effective therapeutics.

Research Experience. Over the past seven years, I have worked as a machine learning researcher in two healthcare startups, where I developed disease screening products using biosignals. I enjoyed every moment of my career, especially since it was satisfying to see how our efforts resulted in products that provide real help to people. Throughout the process of product development, there were many technical challenges, such as domain generalization and noisy labels. Domain generalization (distribution shift problem) was particularly crucial to ensure reliable performance in real-world deployment, so I investigated this problem from multiple perspectives.

One approach I explored was data augmentation, where I tried to simulate demographic distribution shifts using data augmentation. The challenge was how to simulate demographic variations in ECG. To find a hint, I first studied the principles behind ECG signals and found that demographic factors cause differences in heart orientation, position, and chest size, which in turn affect ECG readings. This insight guided me to think of augmentation based on these factors. ECG records the heart's electrical activity from multiple leads, providing different "views" of the heart. Changes in the heart's orientation or position have the same effect as viewing the heart from different angles. Based on concept, I perturbed these views of ECG to simulate changes in heart orientation and position. Specifically, I modeled a graph explaining 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 this approach proved effective on some datasets, since the transformations were applied randomly, rather than targeting specific distributions, it raised questions about its effectiveness in addressing data distribution shifts the model struggles with. Thus, I explored adversarial data augmentation (ADA) as a potential solution, as it generates data distributions that the current model finds challenging. However, I found that existing ADA couldn't simulate temporal changes in ECG. Given the evidence of varying temporal characteristics across different demographic groups, I believed addressing this aspect was crucial to fully tackle 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 traditional time warping. This approach addressed an out-of-distribution issue that existing ADA methods failed to handle, resulting in a 40% improvement in the F1 score on a particular dataset.

In addition, I explored the structural inductive bias of neural networks and its impact on improving generalization to unseen data. Unlike in the computer vision domain, one intriguing observation for ECG data was the superior performance of convolution-based models, which often outperform pre-trained transformers regardless of the size of the data. This led me to investigate using representations of convolution-based models as guides for transformer blocks rather than relying on self-supervised tasks like contrastive learning and masked autoencoders. Specifically, I trained a transformer through a block-by-block knowledge transfer method, where each block of a convolutional neural network guided the corresponding transformer block. This approach enabled the self-attention of the transformer to acquire properties like translational invariance and locality. Additionally, transformers trained with this approach outperformed convolutional networks.

I wondered if transformers could directly exhibit convolutional characteristics (translational invariance and locality) by imposing specific constraints on self-attention without guidance from a pre-trained

convolutional network. By comparing the formulations of self-attention and convolution, I learned that under certain constraints, the attention matrix could function as a depth-wise convolution kernel. Furthermore, the aggregation across attention heads corresponded to point-wise convolution. This perspective allowed for the integration of convolution and self-attention into a single framework. Furthermore, by using tunable coefficients that balance the contributions of self-attention and convolution, the model could effectively acquire properties of both. Experiments across various datasets demonstrated that the proposed method consistently outperformed both naive convolution-based and transformer networks by a considerable margin.

Many of these research efforts were actually applied to enhance our products, and translating research into real-world applications was particularly meaningful to me.

Research Interest. Developing products and conducting research on predictive models are enjoyable and fulfilling. However, a conversation with a doctor last year made me realize that many diseases we can now predict still lack definitive treatments. This left me frustrated, but, at the same time, it fueled my interest in exploring the fundamentals of disease and developing treatments that go beyond prediction. I have particularly developed an interest in bioinformatics.

AI has already had a significant impact on bioinformatics. However, despite these advancements, 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 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, many of these challenges mirror those I faced when applying AI to the medical field, and I would like to use my experience to address these problems.

By tackling these 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 at the gene-to-protein levels through the lens of AI. In relation to this, I hope to explore genomics and transcriptomics, which are essential for identifying key genetic drivers involved in disease. Additionally, I am interested in investigating structural biology, including protein-protein interactions and protein design, as these are important for both understanding diseases and developing effective treatments.

Conclusion. 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. Working as a researcher in such environments also made me realize my deep passion for research. Since I could not fully concentrate on research due to product development demands, I sometimes had to carve out time to pursue my research interests. Last but not least, as a member of a startup, I experienced success. Although it was a mini-success, as the saying goes, "it takes one who has tasted success to believe in success," and I will continue to strive for further achievements.

I see my graduate study as a milestone in my journey to find answers about leveraging AI to liberate humanity from disease. More and more secrets of biology are being unlocked through the power of AI. After graduate studies, with a deeper understanding of bioinformatics and AI, I hope to continue contributing meaningfully to this field, advancing biology and medicine.