Will there ever be a day when humans live without suffering from disease and fully enjoy their lives? Witnessing firsthand a family member suffer from cancer and diabetes gave me an understanding of how illness can limit an individual's physical and mental freedom, deeply motivating me to pursue these questions. With this in mind, I journeyed on a very consistent path, from bioengineering for my master to machine learning researchers in healthcare startups, where I develop diverse product to enhance medical practice. However, I realized that many diseases we can predict still lack definitive treatments. This realization fueled my interest in bioinformatics, which allows for the exploration of the fundamentals of disease and the development of treatments. I hope the PhD studies help me to guide and deepen my research in bioinformatics, serving as a milestone of my research journey.

Research Experience. Over the past seven years, I have worked as a machine learning researcher in two healthcare startups, where I developed diverse 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 real help to people. Throughout the process of product development, there were many technical challenges, such as domain generalization and noisy labels. Domain generalization 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.

These research projects were particularly meaningful, as Many of them have been actually applied to enhance our products.

Research Interest. Developing products and conducting research on predictive models has been enjoyable and fulfilling. However, I realized that many diseases we can now predict still lack definitive treatments during a conversation with medical staffs. This realization 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 based on medical signal. I have particularly developed an interest in bioinformatics.

AI has already has already made 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 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 hope 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 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 effect.

Conclusion. Looking back on my time at two startups, sometimes, the demands of product development often left me little time to solely focus on 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 the academia, where I can fully concentrate on research and study.

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, as a member of a startup, I experienced success. Although it was a mini-success, as the saying that "it takes one who has tasted success to believe in success”, I will continue to strive for further success.

After graduation, with a deeper understanding in bioinformatics and AI, I aspire to make meaningful contributions to advancing biology and medicine, ultimately toward a world where humanity no longer has to suffer from disease.