# Specific Aims:

# Despite decades of work characterizing the mechanisms of splicing, we do not yet understand splicing fully enough to accurately predict the isoform distribution of a sequence across cell types or to describe the mechanisms that will drive the splicing patterns we observe for a new sequence. Deep neural networks (DNN) have shown great promise for predicting splicing from sequence. However, it remains unclear how to synthesize data from highly diverse sources into a single predictive model, how to design new experiments to improve model performance, and how to derive mechanistic understanding from the predictions.

# I propose to address these issues in three complementary aims. In aim 1, I will develop and test strategies for training splicing DNNs across diverse data types and incorporating new data sets as we produce them. In aim 2, I will create strategies for designing massively parallel splicing assays (MPSA) to resolve model uncertainty and evaluate them by performing targeted MPSA experiments. In aim 3, I will use DNN interpretability methods to extract mechanistic hypotheses from splicing models and experimentally test them.

# Aim 1: Develop a foundation model for splicing that integrates data from diverse genomics datasets. As with any scientific effort, building models to reflect our understanding of splicing, requires us to incorporate diverse sources of existing data and continually update the model to reflect new data. I will apply multi-task, transfer, and continual learning techniques to train splicing models across both evolutionary and functional genomics datasets. This will allow information learned in one dataset to transfer to the other datasets. I will apply this technique to improving state-of-the-art splicing models like SpliceAI and Pangolin and trial new model architectures. I will determine how training across datasets, species, and cell types affects model performance.

# Aim 2: Design and perform targeted MPSAs to improve models of splicing. When designing new experiments, it is important to ensure that the results will shed light on uncertainties in your understanding. By estimating model uncertainty as a function of sequence and context we can target new experiments to improve predictions of uncertain sequences in difficult contexts. I will use model ensemble variance to estimate uncertainty and use active learning techniques to identify maximally informative datasets for fine tuning splicing models. I will identify sequences and cell types where model performance is poor or uncertainty is high and use attribution techniques to map sequence features driving poor performance. I will evaluate different model guided library design strategies for improving model performance *in silico*. I will then perform a model guided MPSA and evaluate the effectiveness of active learning techniques with real data. I will evaluate knowledge transfer from the MPSA dataset to the previous datasets focusing on performance in previously uncertain contexts and sequence elements targeted in the MPSA.

# Aim 3: Computationally extract and experimentally test mechanistic hypothesis suggested by splicing models. Predictive ability is necessary but not sufficient for understanding, mechanistic knowledge is also required. DNNs, while predictively powerful, cannot be directly examined for mechanistic insight. I will leverage recent work in DNN interpretability and domain knowledge about splicing to derive mechanistic insights from models. I will identify regulatory regions driving predictions through sequence shuffling. I will follow up with attribution methods such as saliency maps and *in silico* mutagenesis to detect specific motifs that drive splicing behavior. I will use global importance analysis to interrogate the effects of these regions and motifs as well as their position dependence and interactions. I will test these hypotheses by constructing and evaluating variants with systematically varied motif positions and numbers in the lab and by evaluating the effects of treatment with ASOs that block motifs or shRNAs that knock down splicing factors.

# I have formal training in both experimental biology and computer science from my undergraduate studies and my Ph.D. work. This prepares me to carry out the proposed research. This project will allow me to learn cutting edge methods in deep learning and apply that knowledge to outstanding problems in the molecular biology of splicing. The Kinney lab, at Cold Spring Harbor, where I will perform this work has pioneered the development of massively parallel reporter assays as well as machine learning techniques to analyze their results. I will also work closely with the Koo lab which develops deep learning and interpretability methods for genomics data and the Krainer lab which has performed foundational work on the mechanisms of splicing and the development of splice modifying drugs. To gain broader exposure to splicing I will attend the Post-Transcriptional Gene Regulation Gordon Conference and the Eukaryotic mRNA Processing Cold Spring Harbor meeting in alternating years. To gain more training in deep learning I will attend the Biological Data Science and Systems Biology Cold Spring Harbor meetings in the first year and the Biology of Genomes and Probabilistic Modeling in Genomics Cold Spring Harbor meetings in the second year. I will also TA the graduate level Cold Spring Harbor Quantitative Biology course to gain teaching experience.

Together, these research and training opportunities will position me to launch an independent research career focusing on applying massively parallel assays and modeling to understanding deep biological questions in RNA processing.