# Specific Aims:

# Despite decades of work characterizing the mechanisms of splicing, we do not yet have a quantitative understanding of the sequence dependance of splicing. A fully quantitative understanding requires that we can predict the splicing outcomes of an arbitrary sequence in an arbitrary cell type with quantitative accuracy and know the mechanisms responsible for these results. Deep neural networks (DNN) have shown great promise for predicting how sequence governs splicing, however it remains unclear how to synthesize data from highly diverse sources into a single predictive DNN, how to design new experiments to further improve the DNN, and how to mechanistically interpret the DNN.

# I propose to address these three 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 test them by performing targeted MPSA experiments. In aim 3, I will use DNN interpretability methods to extract mechanistic hypothesis from splicing models and experimentally test them.

# Aim 1: Develop a foundation model for splicing that integrates data from diverse genomics datasets. In order to build models of splicing that reflect our full understanding of splicing, we need to be able to incorporate information from diverse sources of existing data and add new information from additional datasets as we produce them. In order to reach this goal, 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 compared to the single task context.

# Aim 2: Design and perform targeted MPSAs to models of splicing. To enable continuous improvement of quantitative splicing models we must be able to estimate model uncertainty as a function of sequence and update the model with new data about uncertain sequences. To address this, I will use variance within model ensembles to estimate model uncertainty and use active learning techniques to identify maximally informative datasets for fine tuning splicing models. I will identify sequences where model performance is poor or uncertainty is high and use attribution techniques to map sequence features driving poor performance. I will design maximally informative MPSA libraries, then simulate results *in silico* using an independent oracle. By fine tuning the models with these datasets, I will evaluate different model guided library design strategies for improving model performance against known ground truth. I will then perform a model guided MPSA and collect real data for fine tuning to evaluate the effectiveness of active learning techniques outside of the simulated context. I will evaluate performance differences between models fine-tuned with unguided MPSA data and model guided MPSA data. I will focus on the ability of the model to transfer learning from the MPSA dataset to the previous datasets, to improve performance in previously difficult contexts, and to resolve uncertainty driven by the sequence elements targeted in the MPSA.

# Aim 3: Computationally extract and experimentally test mechanistic hypothesis suggested by splicing models. For a model to contribute to our understanding of splicing, it needs to provide not just quantitative prediction of outcomes but also mechanistic insight. DNNs while predictively powerful cannot be directly examined for mechanistic insight. I will leverage recent work on DNN interpretability and domain knowledge about splicing to derive mechanistic insights from splicing models. I will identify important regulatory regions driving predictions by randomly shuffling sections of the transcript. I will follow up with attribution methods such as saliency maps and *in silico* mutagenesis to detect motifs that are important drivers of 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 contents in the lab and by evaluating the effects of treatment with ASOs that block important motifs or shRNAs that knock down important splicing factors.

# I have formal training in both experimental biology and computer science from my undergraduate studies and my Ph.D. work. This background prepares me well to carry out the proposed research. This project will allow me to learn cutting edge methods in deep learning and to apply that knowledge to important 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 the machine learning techniques to analyze the results of these experiments. I will work closely with the Koo lab which develops deep learning methods for genomics data and interpretability methods for deriving biological insight from these methods. I will also work closely with the Krainer lab which has performed foundational work on the mechanisms of splicing and the development of splice modifying drugs. As part of these collaborations, I will attend lab meetings and journal clubs and work directly with Adrian Krainer and Peter Koo as well as their lab members for both scientific and career mentorship and advice. To gain broader exposure to splicing I will attend the Post-Transcriptional Gene Regulation Gordon Conference in the first year and the Eukaryotic mRNA Processing Cold Spring Harbor meeting in the second year. To gain more training in deep learning I will attend the Biological data 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 more experience teaching.

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.