# 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.

# Alternative splicing is important for gene regulation; however, it is incompletely understood. Large-scale genomics datasets have been created representing diverse sources of splicing information. These include reference annotations, long and short read RNA-seq, and massively parallel splicing assays (MPSA). However, synthesizing the information from all of these sources into a unified model of splicing is challenging. Deep neural networks (DNN) have demonstrated the ability to extract patterns and synthesize information across diverse datasets. For DNNs to be maximally effective in aiding our understanding of splicing, they must be able to combine information from across datasets, provide information about uncertainty to enable targeting future experiments, learn from new experiments, and be amenable interpretability methods. Current models have failed to meet these requirements.

# I propose to use deep neural networks to extract insight into the biology of splicing from genomics datasets. I will use models to synthesize information across multiple types of splicing datasets, create massively parallel splicing assays optimized for fine tuning model understanding, and develop strategies for extracting insight into the mechanisms of splicing. I will accomplish these goals in 3 complementary aims.

# Aim 1: Train models of splicing across diverse genomics datasets. I will apply multi-task, transfer, and continual learning techniques to train splicing models across multiple 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 diverse datasets including reference isoform annotations, RNA-seq splice junction annotations, and MPSAs across species and cell types affects model performance compared to the single task context.

# Aim 2: Design targeted MPSAs to improve splicing models. I will 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: Extract mechanistic insight from splicing models. I will use interpretability methods to derive mechanistic insights from splicing models. I will use attribution methods such as saliency maps and *in silico* mutagenesis to identify motifs that are important drivers of splicing behavior. I will use global importance analysis to interrogate the effects of these motifs and their interactions. I will investigate knowledge distillation into intrinsically explainable models as a method for creating mechanistic hypotheses.

# My background in both massively parallel assays and computer science position me well to execute this proposal. I have the necessary skills to perform the computational portions of this project from both my undergraduate computer science minor and modeling and bioinformatics projects during my Ph.D. I also have the experience to perform the MPSA from my Ph.D. work performing massively parallel assays and my undergraduate experience majoring in molecular biology and biochemistry and performing wet lab research. However, this project provides an important training opportunity by allowing me to expand my knowledge of cutting-edge deep learning techniques including model interpretability and multi-task, continual, and active learning. Further, this project will help me to explore applications of my skills in the field of splicing. Through this project, I will build expertise at combining deep learning and experiments to extract insight into complex biological mechanisms. While I pursue these scientific goals, I will also be focused on utilizing my position at Cold Spring Harbor lab to gain wider training and mentorship that will help launch my career as an independent researcher. I will attend CSHL meetings and Gordon conferences on RNA processing and quantitative methods, attend lab meetings and journal clubs in the Kinney, Krainer, Koo and McCandlish labs, attend CSHL grant writing and professional development courses, and hone my skill at lecturing through teaching opportunities.

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.