# Specific Aims:

# Alternative splicing is an important factor in gene regulation; however, it is still incompletely understood. A variety of large-scale splicing datasets have been created representing diverse sources of genomics information. These include reference annotations, both long and short read-based RNA-seq datasets, and massively parallel splicing assays (MPSA). However, synthesizing the information from all of these sources into a unified model of splicing has proven challenging. Deep neural networks have demonstrated the ability to extract patterns and synthesize information across diverse datasets. In order to most effectively utilize the power of deep neural networks to aid our understanding of splicing our models need several characteristics. They must be able to combine information from across multiple data sources, they must be able to provide information about uncertainty so that future experiments can be targeted effectively, they must be able to learn from the results of new experiments, and they must provide insight into the underlying mechanisms of splicing. Current models have failed to meet these requirements.

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

# Aim 1: Expand models of splicing to train across diverse genomics datasets. I will create a framework for applying multi-task learning and continual learning techniques to train splicing models across multiple datasets. Training the same model across multiple splicing datasets can allow information learned in one dataset to be transferred to the other datasets. I will apply this framework to extending state-of-the-art splicing models like SpliceAI and Pangolin as well as to testing 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 training in a single task context.

# Aim 2: Develop methods for designing 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 data for model 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 previously learned datasets, to specifically 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 neural network interpretability methods to derive mechanistic insights from splicing models. I will attribution methods such as saliency analysis 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 how these effects interact with each other. 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 will 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 necessary skills to perform the MPSA I propose 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 continual learning, active learning, and model interpretability. Through this project, I hope to build skills that will help me combine deep learning and experiments to extract insight into complex mechanisms. Further, this project will allow me to learn about the field of splicing and how to apply my expertise in massively parallel assays and computational modeling to this area. 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.

1. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2015).

2. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

3. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).

4. Liao, S. E., Sudarshan, M. & Regev, O. Machine learning for discovery: deciphering RNA splicing logic. *bioRxiv* 2022.10.01.510472 (2022) doi:10.1101/2022.10.01.510472.

5. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).

6. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

7. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

8. Wong, M. S., Kinney, J. B. & Krainer, A. R. Quantitative Activity Profile and Context Dependence of All Human 5’ Splice Sites. *Mol. Cell* **71**, 1012-1026.e3 (2018).

9. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).

10. Adamson, S. I., Zhan, L. & Graveley, B. R. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol.* **19**, 71 (2018).

11. Soemedi, R. *et al.* Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* **49**, 848–855 (2017).

12. Cheung, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol. Cell* **73**, 183-194.e8 (2019).

13. Baeza-Centurion, P., Miñana, B., Valcárcel, J. & Lehner, B. Mutations primarily alter the inclusion of alternatively spliced exons. *Elife* **9**, (2020).

14. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* **176**, 549-563.e23 (2019).