# Specific Aims:

# Alternative splicing is an important factor in gene regulation; however, its regulation is still incompletely understood. Mutations that change splicing patterns cause diverse genetic diseases.1 Current models of splicing regulation are informed by large scale splicing datasets.2–4 These datasets include transcript isoform reference annotations like those found in GENCODE,5 RNA-seq reads covering splicing junctions in different contexts like those found in GTEx and ENCODE,6,7 and massively parallel splicing assays (MPSAs) which collect information on splicing outcomes for particular variants.4,8–14 However, current practice does not efficiently utilize this data to train splicing models because models are not directly trained in a cross-dataset fashion and therefore information learned from one dataset cannot inform learning on another. Further, new data collection is not targeted to maximally improve model performance. These factors combine to limit our ability to predict splicing outcomes.

# I propose to develop methods for improving splicing models by training them across multiple types of datasets and to design experiments to maximally improve model performance. I will use these methods both in the context of improving current state-of-the-art splicing models and for training new model architectures. 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 trialing 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 model performance. I will identify sequences where model performance is poor or uncertainty is high and use model attribution techniques to map sequence features driving poor performance. I will design maximally informative MPSA libraries, then simulate results *in silico*. By fine tuning the models with these datasets, I will evaluate different model guided library design strategies for improving model performance.

# Aim 3: Generate new MPSA data and evaluate effects on model understanding. I will perform a full model guided MPSA and collect data for model fine tuning to evaluate the effectiveness of these 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 and to specifically improve performance on previously poor contexts and to resolve uncertainty driven by the targeted sequence elements.

# My background positions me well to execute this proposal, yet this proposal offers a training opportunity that will help me found an independent lab. In my doctorate I used massively parallel assays and computational modeling in microbiology, protein engineering, and CRISPR tool development providing me with the foundational skills I will need for this proposal. This proposal will allow me to explore new applications of these skills in a field with opportunities to do fundamental science with applications in human health. Specifically, this project will provide training in splicing and in more advanced deep learning techniques. 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).