**Significance**

Background

Alternative splicing is a fundamental aspect of eukaryotic gene regulation and more than 90% of human genes are alternatively spliced.1 Splicing is the process of identifying introns and exons in a pre-mRNA sequence and removing the introns while stitching together the exons to form an mRNA. Errors in this process are involved in a wide variety of diseases including familial dysautonomia, early onset Parkinson disease, and cancer.2 Further, recent efforts have produced drugs that modulate splicing outcomes to treat spinal muscular atrophy, Huntington's disease, Duchenne muscular dystrophy, and cancer.3 There are four essential components that mark a potential new splice junction these are the 5’ splice site, branch point, polypyrimidine tract, and 3’ splice site. The branch point and polypyrimidine tract are contained in the intron while the 5’ and 3’ splice sites mark the boundaries between the intron and the 5’ and 3’ exons respectively. Splicing is initiated by binding of the U1 snRNP to the 5’ splice site, binding of SF1 to the branch point, and binding of U2AF to the polypyrimidine tract and 3’ splice site, these factors interact to form the E complex which then initiates splicing. In humans, the initial interaction of these factors frequently occurs across the exon in a process called exon-definition, but the interaction then transfers to cross the intron that will be removed before splicing occurs. This complex then recruits additional spliceosome factors and undergoes a series of rearrangements to yield an active complex. In this process, the U2 snRNP replaces SF1 and U2AF to bind the branch point and U1 is replaced by the U6 snRNP. The actual splicing reaction then occurs as a sequence of two nucleophilic attacks, severing first the 5’ exon and then the 3’ exon from the intron and sealing the two exons together.4 In humans, however while the strength of a splice junction can be controlled by adjusting the affinity of these regulatory sequences for their respective spliceosomal factors, much finer regulation of splicing is required to allow dynamic changes to isoform populations. This is achieved through the binding of proteins called splicing factors to motifs called splicing enhancers or suppressors that may be present in the introns or the exons of the pre-mRNA. These factors affect the recruitment of the spliceosome components and allow specific isoform changes to be tightly regulated through changes in splicing factor abundance or activity.5 Additionally, RNA secondary structure changes can dramatically affect which isoforms are produced.6,7 How these influences integrate to produce the final splicing outcome is only poorly understood. A major goal of the splicing field has been to understand splicing fully enough to accurately predict the isoform distribution of a sequence across cell types and to describe the mechanisms that will drive the splicing patterns we observe for a new sequence. However, this has proven challenging.

DNNs have revolutionized modeling efforts across a wide array of fields and have shown great success at approximating unknown functions and extracting complex relationships from data. However, this performance comes at the cost of encoding the decision making process of the model in a form that cannot be directly understood.

The three most common types of DNN architectures are feed forward networks or multilayer perceptrons, convolutional neural networks, and transformers. Sufficiently large feed forward networks can approximate arbitrary functions

Deep neural networks (DNN) have shown great promise for predicting splicing from sequence. Models such as spliceAI8 and Pangolin9 have demonstrated the suitability of

1. Deep learning background
   1. What is a DNN and how do they work
   2. Multi-task, transfer, and continuous learning
      1. Allow transfer of information across datasets
      2. Can improve performance on all tasks
      3. Differences between the strategies
   3. Active learning
      1. Allow efficient improvement of existing models
      2. Requires estimates of uncertainty
   4. Interpretability
      1. Goal is to extract understanding from predictive models
      2. Post hoc methods
         1. Occlusion
         2. Saliency/ISM
         3. GIA
      3. Interpretability by design (maybe)
      4. Pruning and distillation (maybe)

Massively parallel splicing assays (MPSAs) offer the potential to rapidly profile the regulatory mechanisms of splicing in genes like PKM. These assays use high-throughput sequencing to measure quantitative splicing phenotypes for 10s of thousands of variants in parallel.10–24 They work by linking the sequence of the pre-mRNA to a barcode and then either compare barcode abundances between two conditions as a proxy for splicing10,16,19–21,23–25 or directly sequence the splice junction of interest and the barcode to observe the isoform distribution of each variant.12,14–16,18,22 These assays have allowed identification of human sequence variants that cause splicing changes,20 in depth study of how drugs interreact with splicing regulation,24 and detailed mechanistic studies on splicing regulations.10,15,16,19,21 However, MPSAs have been held back by two barriers to progress.

Barrier to progress 1: …

Barrier to progress 2: …

Barrier to progress 3: …

*This proposal will overcome these barriers by …*

THE KEY ADVANCE FOR EACH SECTION (underline the actual key advance in each paragraph)

We propose pursuing these goals through 3 complementary aims. Aim 1…. Aim 2 …. Finally, aim 3 ...

Innovation

Our work will enable broader use of MPSAs to interrogate the mechanisms of splicing by extending the technique to work in the context of complex isoform distributions and correlated splicing outcomes. It will further enable wider use by providing end to end analysis and modeling software for MPSA data. This work is driven by innovations that will change the practices of the MPSA field and the broader splicing community.

**Innovation 1: …** intro **…** **Aim 1** **…** and **aim 2** **…**

**Innovation 2: …** i

**Innovation 3: …** i

**Approach**

**Aim 1: Develop a foundation model for splicing that integrates data from diverse genomics datasets.**

Goal: ….

Approach: ….

Pitfalls and alternative approaches: ….

**Aim 2: Design and perform targeted MPSAs to improve models of splicing.**

Goal: ….

Approach: ….

Pitfalls and alternative approaches: ….

**Aim 3: Computationally extract and experimentally test mechanistic hypothesis suggested by splicing models.**

Goal: ….

Approach: ….

Pitfalls and alternative approaches: ….

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

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

3. Neil, C. R. *et al.* Reprogramming RNA processing: an emerging therapeutic landscape. *Trends Pharmacol. Sci.* **43**, 437–454 (2022).

4. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).

5. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).

6. Jin, Y., Dong, H., Shi, Y. & Bian, L. Mutually exclusive alternative splicing of pre‐mRNAs. *Wiley Interdiscip. Rev. RNA* **9**, e1468 (2018).

7. Xu, B., Meng, Y. & Jin, Y. RNA structures in alternative splicing and back-splicing. *Wiley Interdiscip. Rev. RNA* **12**, e1626 (2021).

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

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

10. Ke, S. *et al.* Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).

11. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 1–8 (2016).

12. 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).

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

14. Cortés-López, M. *et al.* High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling CD19 splicing and CART-19 therapy resistance. *Nat. Commun.* **13**, 1–17 (2022).

15. Schirman, D., Yakhini, Z., Pilpel, Y. & Dahan, O. A broad analysis of splicing regulation in yeast using a large library of synthetic introns. *PLoS Genet.* **17**, e1009805 (2021).

16. Mikl, M., Hamburg, A., Pilpel, Y. & Segal, E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat. Commun.* **10**, 1–14 (2019).

17. Braun, S. *et al.* Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat. Commun.* **9**, 1–18 (2018).

18. Souček, P. *et al.* High-throughput analysis revealed mutations’ diverging effects on SMN1 exon 7 splicing. *RNA Biol.* **16**, 1364–1376 (2019).

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

20. 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).

21. 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).

22. 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).

23. 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).

24. Ishigami, Y. *et al.* Specificity, synergy, and mechanisms of splice-modifying drugs. *bioRxiv* 2022.12.30.522303 (2022) doi:10.1101/2022.12.30.522303.

25. North, K. *et al.* Synthetic introns enable splicing factor mutation-dependent targeting of cancer cells. *Nat. Biotechnol.* **40**, 1103–1113 (2022).