A099 · Folding It In: Structure-Aware Deep Splicing Models

Presenter: Utkarsh Goel — Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Authors: Utkarsh Goel, Arush Ramteke, Oded Regev

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Deep learning models such as SpliceAI and AlphaGenome predict splicing outcomes from sequence alone with remarkable accuracy. However, recent work has shown that these state-of-the-art models are blind to RNA structure, failing to respond to structure-altering mutations, including clinically relevant variants.

Here, we ask whether this failure mode can be mitigated by explicitly providing RNA secondary structure to these models. We supply the probabilities that windows of 1–8 nt ending at each nucleotide are unpaired as eight additional input channels to OpenSpliceAI and train an otherwise identical sequence-only model as a control.

On a held-out genomic test set, structure improved overall accuracy, reducing errors by ~9%, while retaining an essentially identical parameter count. On previously established tests of structural blindness, our structure-aware model substantially outperforms comparable sequence-only models: (1) On a held-out synthetic dataset, prediction error grows steeply with exon folding stability across all sequence-only models; our structure-aware model improves this by nearly two-fold. (2) On compensatory mutation series that disrupt and then restore a stem loop, only our structure-aware model demonstrates sensitivity to the effects of structure on exon inclusion: no sequence-only model reproduces the full trajectory, even directionally.

Overall, we show that explicit RNA secondary structure captures regulatory features that are not reliably learned from sequence alone, supporting biologically motivated inductive biases as a practical route to more robust genomic deep learning models. To the best of our knowledge, our results also provide the first evidence for a broad role of RNA structure in genomic splicing.

Folding It In: Structure-Aware Deep Splicing Models

Utkarsh Goel[1]*, Arush Ramteke[1], Oded Regev[1]
[1]*Courant Institute, New York University*

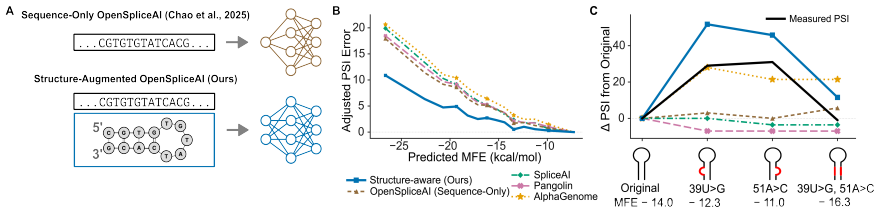

Figure 1. ***(A)*** *State-of-the-art models use RNA sequence alone; our model adds predicted RNA structural features as input.* ***(B)*** *Prediction error (predicted minus measured percent spliced in (PSI)) for sequences binned by predicted minimum free energy (MFE), adjusted for measured inclusion following [5], with each model shifted so its least-folded bin reads zero; a steeper slope indicates stronger structure bias.* ***(C)*** *Measured and predicted change in PSI across a series of splice-affecting mutations (red) that disrupt an exonic stem loop (black) and then restore it with compensatory mutations.*

Introduction. RNA splicing is a fundamental step in gene regulation, governed by core splice sites, splicing enhancers and silencers. Deep learning models such as SpliceAI, Pangolin and AlphaGenome predict splicing outcomes from sequence alone with remarkable accuracy [1–4].

Splicing is also modulated by RNA structure [6]. However, recent work showed that state-of-the-art splicing models systematically miss these effects: prediction errors increase with RNA structural stability, and models fail to identify clinically relevant splice-altering variants that act through structure [5].

Here, we ask whether splicing prediction can be improved by directly incorporating predicted RNA secondary structure as additional input to such models.

Methods. We represent RNA secondary structure using the probability that windows of 1–8 nt ending at each nucleotide are unpaired, as predicted by RNAplfold [7]. We train OpenSpliceAI with these eight probabilities as additional input channels alongside the one-hot encoded sequence [3] (Fig. 1A). We train on MANE protein-coding transcripts using the same train-test split as in the original SpliceAI work and five random seeds [4]. For comparison, we also re-train the same architecture without providing structure channels.

Structure shapes splicing decisions genome-wide. On held-out genomic splice sites, the structure-aware model consistently outperforms its sequence-only control: accuracy $95.8 \pm 1.0\%$ versus $95.4 \pm 1.2\%$ (a nearly 9% reduction in error rate), and top-k accuracy $94.1 \pm 0.9\%$ versus $93.9 \pm 0.9\%$ (mean $\pm$ SD across five random seeds). Importantly, this improvement is achieved with the exact same architecture and essentially the same number of parameters.

Structure channels reduce structure-associated prediction error. Liu et al. show that sequence-only models systematically over-predict exon inclusion for more stably folded sequences [5]. We repeat this analysis on the same held-out MPRA dataset, regressing prediction error against predicted minimum free energy (MFE) while controlling for exon inclusion [5]. All tested sequence-only models reproduce the strong dependence of error on MFE, with slopes from -1.07 to -0.93, despite differences in architecture and training data. In contrast, the structure-aware model shows a substantially weaker dependence with a slope of -0.55 (Fig. 1B).

Structure-aware predictions capture structure-driven splicing. Liu et al. tested structure sensitivity using multiple series of compensatory mutations that disrupt and then restore a stem loop [5,6]. Either single mutation breaks the stem and strongly increases exon inclusion, while the double mutant restores base-pairing and returns inclusion close to the original, making RNA structure the dominant signal across the series. Sequence-only models fail to capture the effects of one or both mutations; only the structure-aware model follows this trajectory (Fig. 1C). Similar results hold across four additional series.

Conclusion. Explicit secondary structure information captures regulatory features that are not reliably learned from sequence alone, suggesting a practical route to more robust splicing prediction models. To the best of our knowledge, our results also provide the first evidence for a broad role of RNA structure in genomic splicing; previously, most evidence was anecdotal, from a few specific exons.

References.

1. Avsec, Ž. *et al. Nature* **649**, 1206 (2026).
2. Zeng, T. & Li, Y. I. *Genome Biol.* **23**, 103 (2022).
3. Chao, K.-H. *et al. eLife* **14**, RP107454 (2025).
4. Jaganathan, K. *et al. Cell* **176**, 535 (2019).
5. Liu, S. *et al. Genome Biol.* (2026); *bioRxiv* 2025.11.28.691252.
6. Liao, S. E. *et al. Proc. Natl. Acad. Sci.* **120**, e2221165120 (2023).
7. Bernhart, S. H. *et al. Bioinformatics* **22**, 614 (2006).