Computational Biomedicine
https://doi.org/10.70401/cbm.2026.0019

Supplementary information

Isoform function prediction via knowledge distillation from alternative splicing

Tong Gu, Jun Wang

1.Implementation Details

The SpliceEM framework was implemented in PyTorch. The model was optimized using AdamW with a learning rate of 2×10^{-4} to ensure stable convergence across multi-modal features. The hidden dimension was set to 768 to align with the output dimension of the pre-trained language models used for GO term semantic embeddings. The heterogeneous graph transformer (HGT) backbone was configured with two message-passing layers to capture multi-hop topological dependencies without introducing over-smoothing. To model different semantic features, the sequence-related and GO-related attention modules utilized 8 attention heads, while each HGT layer utilized 4 attention heads. Within the self-distillation framework, the EMA teacher model was updated using a decay coefficient of 0.99 to maintain a stable historical consensus. Finally, for the loss balancing schedule, the distillation weight was dynamically adjusted via a Gaussian ramp-up strategy following an initial burn-in stage to introduce soft labels steadily during early training.

2.Calibration of Predicted Probabilities

While conventional metrics such as AUPRC, AUROC, Fmax, and Smin are essential for addressing extreme class imbalance, they primarily evaluate ranking quality and threshold-based classification behavior without directly assessing the reliability of the predicted probabilities. To rigorously validate whether the model's prediction confidence faithfully aligns with the true likelihood of a positive functional annotation, we conducted a comprehensive probability calibration analysis on the test set (**Figure S1**).

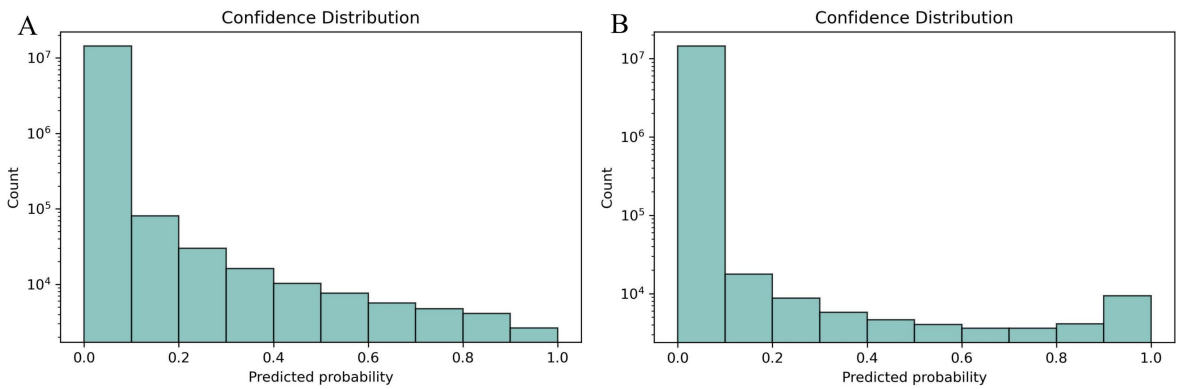

Figure S1. Confidence distributions of the predicted probabilities. The histograms (shown in log scale) illustrate the distribution of predicted probabilities before and after *post-hoc* temperature scaling, reflecting the model's confidence across the test set.

Specifically, we quantified model calibration using the expected calibration error (ECE), which measures the weighted average discrepancy between predicted confidence and empirical accuracy, alongside the Brier score, which provides a strict penalty-based assessment of overall probabilistic accuracy. Evaluation of the raw probability outputs from the optimized model yielded a highly competitive baseline ECE of 0.0070 and a Brier score of 0.00152.

To further minimize overconfidence and refine probabilistic reliability, we implemented a standard *post-hoc* temperature scaling procedure. A single scalar temperature parameter was iteratively fitted on the validation set to rescale the output logits without altering the original prediction rankings, and was subsequently evaluated on the independent test set. Following this calibration step, the test-set ECE was significantly reduced to 0.00222, with the Brier score further decreasing to 0.00137. These results demonstrate that the predicted probabilities align with the observed functional outcomes, supporting the validity of the subsequent biological analyses.

3.Additional Splicing-Regulatory Patterns

To further validate the predictive capacity of SpliceEM across diverse biological scenarios, we analyzed two additional regulatory patterns beyond the antagonistic switch discussed in the main text.

First, we examined the GNAS gene (ENSG00000087460) as a representative of the dominant isoform pattern. As shown in **Figure S2**A, the primary transcript (ENST00000371095) maintains a dominant expression profile across skin, kidney, and cell line environments. It carries essential functions related to the regulation of transcription by RNA polymerase II (GO:0006357). The alternative transcript (ENST00000480975) remains close to background expression levels across these contexts. This confirms that certain genes rely on a single dominant isoform rather than expression switching to maintain core biological stability.

Second, we investigated the DLG1 gene (ENSG00000075711) to illustrate single-context emergence. As shown in **Figure S2**B, one isoform (ENST00000659716) is relatively expressed in skin and kidney tissues to support DNA-templated transcription regulation (GO:0045893). The alternative isoform (ENST00000412364) emerges predominantly in cell lines. This emergence is accompanied by a functional shift toward cell-substrate junction

organization (GO:0150115) and cell-cell adhesion (GO:0098609). This pattern demonstrates that alternative isoforms can be specifically upregulated to meet specialized requirements in distinct cellular environments without globally replacing the primary transcript.

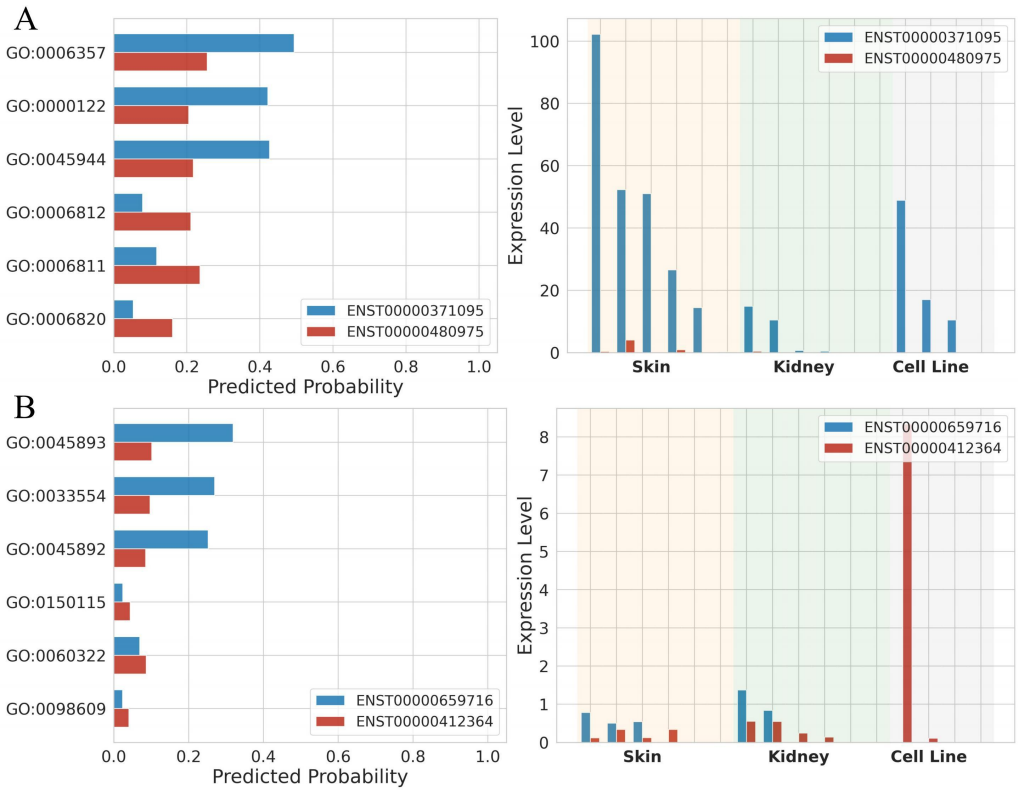

Figure S2. Additional splicing-regulatory patterns identified by SpliceEM. (A) Functional predictions and cross-sample expression profiles for the GNAS gene. This illustrates a dominant isoform pattern where one transcript maintains higher expression across all tested contexts; (B) Functional predictions and cross-sample expression profiles for the DLG1 gene. This demonstrates single-context emergence where a specific isoform is upregulated in a distinct environment to support specialized functions.