Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable machine learning enables de novo mapping of cell type-specific RNA splicing regulation from scRNA-seq data

Xuerui Yang

yangxuerui@tsinghua.edu.cn

Tsinghua University https://orcid.org/0000-0002-7731-2147

Xianke Xiang

Chongqing University

Article

Keywords: single-cell RNA sequencing, alternative splicing, splicing regulation, RNA-binding protein (RBP), ensemble learning, interpretable machine learning

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9148958/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on September 8th, 2026. See the published version at https://doi.org/10.1038/s41467-026-77627-5.

Interpretable machine learning enables *de novo* mapping of cell type-specific RNA splicing regulation from scRNA-seq data

Xianke Xiang[1,2], Xuerui Yang[2*]

[1] Institute of Hepatopancreatobiliary Surgery, Chongqing General Hospital, School of Medicine, Chongqing University, Chongqing, 401147, China

[2] MOE Key Laboratory of Bioinformatics, Center for Synthetic & Systems Biology, School of Life Sciences, Tsinghua University, Beijing, China

*Correspondence: Xuerui Yang, School of Life Sciences, Tsinghua University, Beijing 100084, China. Tel: 86-10-62783943. Email: yangxuerui@tsinghua.edu.cn

Abstract

Context-dependent regulation of alternative splicing (AS), largely mediated by RNA-binding proteins (RBPs), is a key post-transcriptional mechanism shaping diverse biological processes. Experimental approaches for probing splicing regulation, such as CLIP-based assays and RBP perturbations, suffer from low throughput, poor physiological relevance, and bulk resolution that overlooks cellular heterogeneity. Here, we present CASREL (cell-specific alternative splicing regulation inference via explainable learning), a machine learning framework that reconstructs RBP-AS regulatory circuitry directly from single-cell RNA sequencing data without reliance on prior RBP-RNA binding annotations. CASREL integrates ensemble learning with SHAP-based interpretation to uncover RBP-AS associations, leveraging unique features of single-cell splicing profiles, including polarized isoform usage, minimal averaging of regulatory programs, and resilience to expression noise. Applications across diverse tissues and cell types demonstrate its robustness, accuracy, and biological relevance, establishing CASREL as a powerful method for *de novo* mapping of cell-specific RNA splicing regulation in physiological contexts.

Key words: single-cell RNA sequencing, alternative splicing, splicing regulation, RNA-binding protein (RBP), ensemble learning, interpretable machine learning

Introduction

The sophisticatedly regulated alternative splicing (AS) is a fundamental mechanism that expands transcriptomic and proteomic diversity in eukaryotes by enabling context-dependent plasticity of gene expression [1]. Over 90% of human multi-exon genes undergo AS [2]. By selectively utilizing splice sites, AS modulates protein structure [3], localization [4, 5], stability, and function [6], thereby playing critical roles in physiological processes such as cell differentiation and tissue development. Dysregulation of AS has been widely associated with numerous diseases, including cancer, neurological diseases, and metabolic disorders [1, 7, 8].

Systematically dissecting the context-specific splicing regulation programs remains a major challenge [9]. Splicing regulation is inherently complex, being shaped by both cellular contexts and physiological conditions [10]. AS patterns arise from intricate interactions between cis-regulatory RNA elements and trans-acting factors, particularly RNA-binding proteins (RBPs) and

other associated cofactors [11]. The regulatory mechanisms vary substantially across cell types, developmental stages, and environmental conditions [12-14], highlighting the needs for systematic strategies to decode context-dependent, cell-specific AS regulation programs.

Experimental methods have provided valuable insights into RBP-mediated AS regulation. CLIP-based assays map RBP-RNA interactions, but their application is limited by availability of high-quality RBP antibodies, lack of information on regulatory directionality, and high cell-input requirements [15]. Perturbation-based AS profiling reveal changes in splicing patterns upon RBP knockdown, but with difficulties in differentiating direct and indirect splicing regulations. Both approaches are low-throughput and typically restricted to a few preselected RBPs in limited model cell lines, making them unsuitable for dissecting global AS regulatory programs in physiological contexts. Moreover, these bulk cell-based assays lack the resolution to resolve regulatory circuits in specific cell subpopulations, thereby overlooking both cellular heterogeneity of AS regulation in complex tissues and its dynamic shifts during tissue development or disease progression.

Single-cell RNA sequencing (scRNA-seq) offers transformative opportunities by resolving splicing heterogeneity at cellular resolution [16, 17]. Unlike bulk RNA-seq, where mixed isoforms obscure cell-specific AS programs, single-cell data often exhibit polarized splicing distributions, with transcripts from a gene dominated by a single isoform [16, 17]. Such polarized AS patterns likely reflect dominant, cell-specific regulatory programs that underpin finely tuned tissue development and specialized cellular functions [18]. Moreover, because single-cell AS profiles are based on isoform ratios, they are inherently resistant to gene expression fluctuations due to technical variations and transcription bursts. Together, these properties make single-cell AS landscapes a clean and highly informative resource for inferring splicing regulation [17].

Here, we present CASREL (cell-specific alternative splicing regulation inference via explainable learning), a machine-learning framework for reconstructing RBP-AS regulatory circuitry directly from scRNA-seq data. CASREL combines ensemble learning to predict single-cell AS profiles from RBP expression with SHAP-based model interpretation to extract the most influential RBP-AS relationships. Unlike conventional approaches that rely on prior knowledges such as RBP-binding motifs or CLIP-seq annotations of RBP-RNA interactions, CASREL leverages the unique advantages of single-cell splicing data for *de novo*, cell type-specific reconstruction of RBP-AS regulatory circuitry within complex tissues, which is unattainable with bulk RNA-seq. Applications across diverse scRNA-seq datasets from tumors, immune cells, and brain cells demonstrate that CASREL

not only predicts splicing patterns with high accuracy, but also uncovers biologically relevant RBP-AS relationships, many supported by RBP perturbation experiments. Together, these results establish CASREL as a robust framework for elucidating the dynamic and context-dependent control of alternative splicing under physiological conditions.

Results

The CASREL framework for dissecting RNA splicing regulation from scRNA-seq data

Compared with bulk RNA-seq, single-cell RNA-seq provides unique advantages for mining of RNA splicing regulation circuits, including (1) large sample sizes at single-cell resolution, (2) pronounced cell heterogeneity producing high statistical power, and (3) polarized AS profiles reflecting dominant, cell-specific splicing programs. CASREL is designed to leverage these properties to identify cell type-specific regulators of individual AS events, with a focus on RBPs.

CASREL adopts an interpretable machine learning strategy with a three-phases workflow: data preprocessing, AS modeling, and model interpretation (Fig. 1). (i) Data preprocessing. AS events are first identified *de novo* and quantified as splicing probabilities per cell from scRNA-seq data (typically SMART-seq), with our previous single-cell AS clustering approach SCASL [17]. Given the polarization of single-cell AS probabilities [17, 19], these values are discretized into 4 categories: non-spliced (splicing probability < 0.4), neutral (0.4-0.6), spliced (> 0.6), and NA (missing coverage) (Fig. 1). Such categorization reduces technical noise from sparse read counts. It explicitly treats missing AS probability data as a separate class, reflecting potential biological relevance of single cell-specific transcription stochasticity [2, 20].

(ii) AS modeling. For each AS event, CASREL trains a predictive model using RBP expression as input features and categorical AS profiles as outputs (Fig. 1). We employ gradient boosting decision trees (GBDT) implemented in LightGBM [21], an ensemble learning model well suited for sparse, imbalanced, and high-dimensional data [22-24], which also offers strong interpretability [25]. The LightGBM framework allows efficient selections of information-rich samples and features through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) (Methods for details). In addition, LightGBM natively supports categorical AS values, avoiding the sparsity issues of one-hot encoding [21]. Model robustness is further ensured by *k*-fold cross-validation, where mutually exclusive training and test datasets are used to train *k*

independent GBDT models (Fig. 1). Consensus across all these models prevents information leakage, reduces overfitting, and enhances generalization to unseen data.

(iii) Model interpretation. The GBDT models trained above are interpreted using SHAP (SHapley Additive exPlanations) [26], a game theory-based framework that attributes predictive outcomes to individual RBPs (Fig. 1). This framework generates a three-dimensional SHAP matrix capturing the contribution of each RBP to each AS category for each single-cell sample. By aggregating SHAP values within particular cell types, CASREL quantitatively identifies high-impact RBP-AS associations, which constitute the cell type-specific splicing regulatory circuitry (See Methods for details).

Figure 1

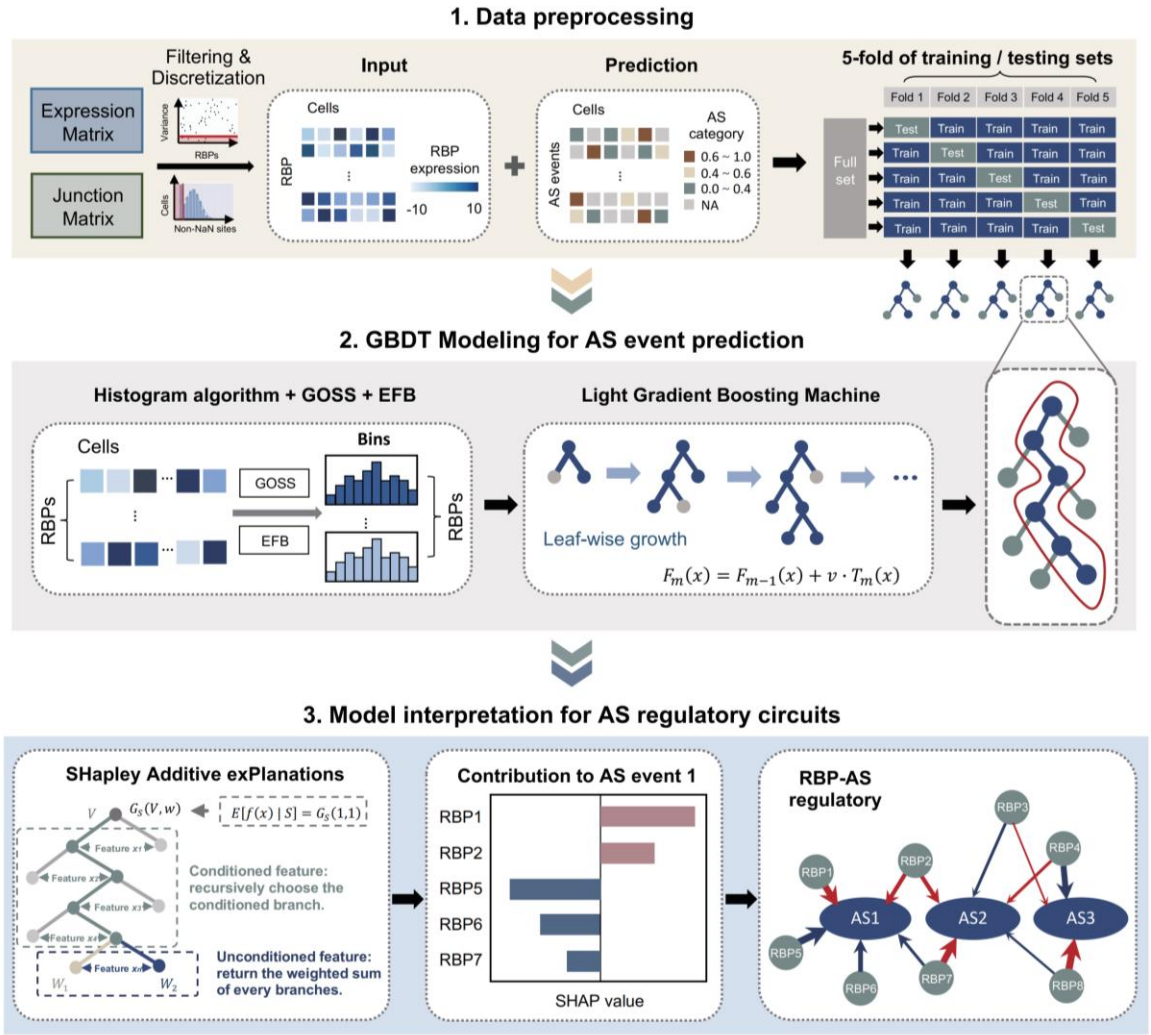


Figure 1. Overview of CASREL.

Schematic of CASREL, which takes single-cell RBP expression profiles and AS profiles as input to reconstruct RBP-AS regulatory circuitry. The pipeline is composed of 3 major steps: 1) Data preprocessing, 2) Modeling of single-cell AS profiles, and 3) Model interpretation to infer RBP-AS regulations.

Benchmarking the CASREL pipeline built on different models

A key component of CASREL is the machine learning model used to fit AS profiles with RBP expressions (Fig. 1). Here, we benchmarked the CASREL pipelines built on three widely used models: logistic regression (LR), a fully connected neural network (FCN), and gradient boosting decision trees (GBDT). First, using identical datasets and training schemes, we evaluated the AS prediction accuracies of the three models on 4 datasets, including two TNBC tumor cell datasets (TNBC-1 and TNBC-2), hepatocellular carcinoma (HCC) tumor immune cells, and mouse brain cells (from Smart-seq2 for the first three, 10X for the last). Each dataset was split into 80% training and 20% testing. Predictions of categorized AS profiles with normalized RBP expressions were compared to ground-truth. Both GBDT and FCN substantially outperformed LR, with GBDT achieving the highest accuracy in several datasets (Fig. S1A), indicating limited capability of the simple logistic regression model to capture complex nonlinear RBP-AS relationships.

We next assessed the interpretability of the three models, which is critical for inferring splicing regulators. LR coefficients approximate RBP-AS associations in the LR model, while SHAP values quantify contributions of individual RBPs to AS profiles in FCN and GBDT. Across all 4 datasets, only GBDT produced SHAP values with a long-tail distribution, indicating that a small subset of RBPs exert strong influence in shaping the AS landscapes (Fig. S1B).

We further compared the specific RBP-AS associations inferred by the three models. For example, CD44s, generated by exon 6-14 skipping in CD44, is a prognostic marker for malignancy and regulates cell adhesion and migration [27]. From the TNBC-1 dataset, LR and FCN inferred both promotive and inhibitory effects of the same RBPs on the formation of CD44s (Fig. 2A). This pattern was not unique to CD44s. Across all four datasets, over 80% of RBPs were inferred to exert contradictory effects on individual AS events (Fig. 2B), which is biologically implausible. For LR, this likely stems from its inability to capture nonlinear, multi-plex regulation of RNA splicing.

For FCN, this may result from the limited interpretability and potential overfitting of a complex model with small-scale data.

By contrast, GBDT inferred predominantly monotonic effects for the majority of RBPs on specific AS events (Fig. 2B). For instance, the CASREL pipeline based on GBDT identified mutually exclusive RBPs that either promoted or inhibited the formation of CD44s (Fig. 2A). Among them, ESRP1 (repressor) [28], HNRNPL (activator) [29], and HNRNPM (activator) [30] (Fig. 2A) have been previously established as potent regulators of CD44s in breast cancer cells, thereby regulating epithelial-mesenchymal transition (EMT).

Based on the analyses above, we adopted GBDT as the core model for CASREL. With k-fold cross-validation (k=5), all models showed high accuracy in predicting AS (CV < 0.5% for AS categories, Fig. S2A), and the SHAP values were also highly consistent (Fig. S2B), underscoring CASREL's robustness. Finally, we assessed the performance of CASREL across varying cell numbers by down-sampling the original datasets. CASREL maintained robust performance with hundreds of cells (>400 cells recommended), as reflected by both AS prediction accuracy (Fig. S3A) and monotonic RBP effects on specific AS events (Fig. S3B). Importantly, most RBP-AS regulatory circuits inferred from as few as 400 cells were preserved with larger cell numbers (Fig. S3C), indicating that CRASEL reliably identify AS regulators from relatively small datasets. This underscores its utility for resolving cell- or subtype-specific splicing regulation programs when sample sizes are limited, though the minimal cell number requirement may vary across datasets.

Figure 2

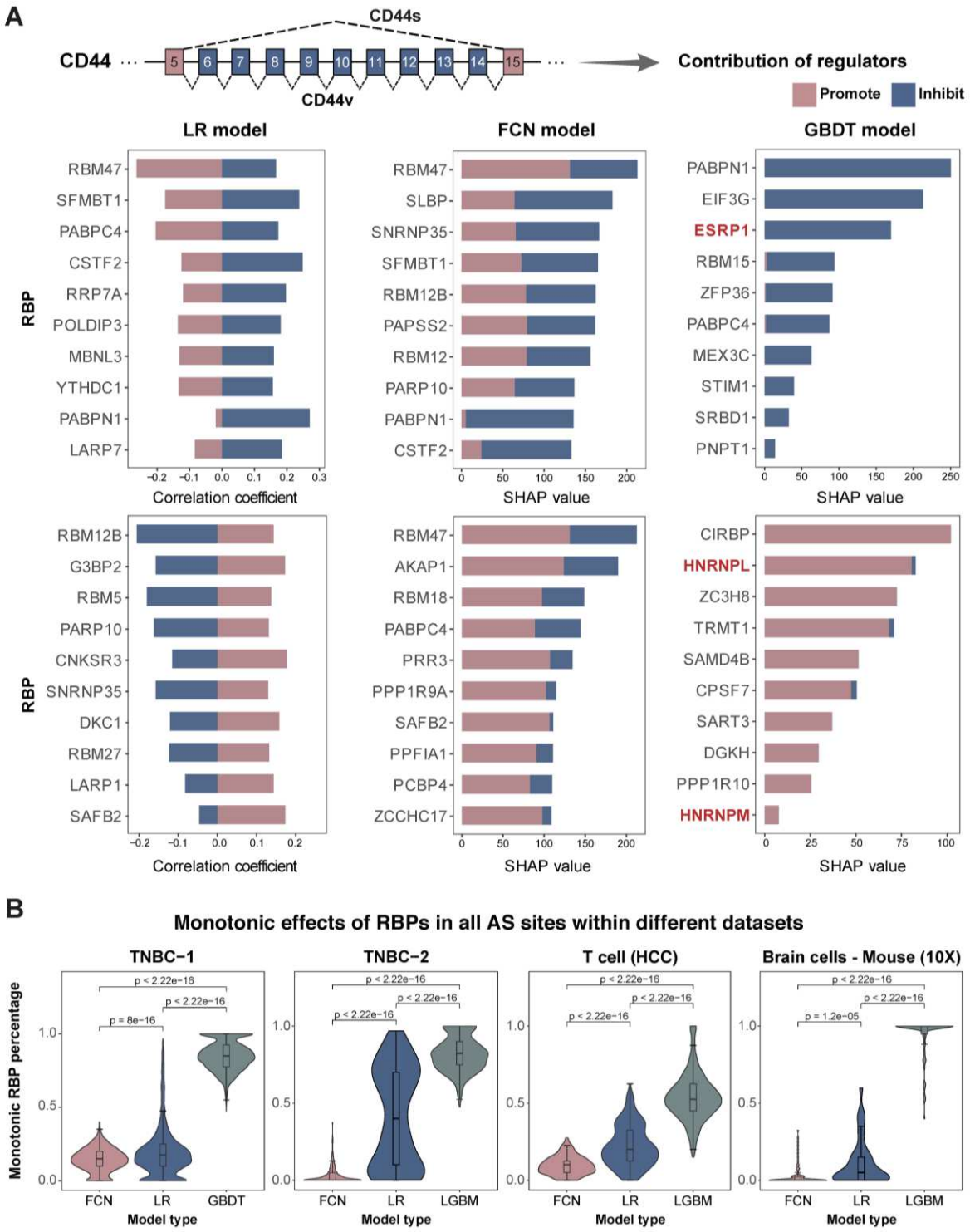


Figure 2. Evaluation of model interpretability for inference of splicing regulators.

(A) Contributions of regulatory RBPs to CD44 exon 6–14 skipping across three models: linear regression (left), FCN (middle), and GBDT (right). Positive (red) and negative (blue) values represent promotion or inhibition of exon skipping, respectively. RBPs are ranked by contribution magnitude. Top and bottom panels display the ten strongest RBPs promoting exon inclusion or exon skipping. Literature-validated CD44s regulators are highlighted in red.

(B) Percentages of RBPs exhibiting monotonic regulatory effects on AS events as predicted by CASREL based on different models applied on four datasets. Violin plots with box plots show the percentage distributions for all AS events. Statistical significance was determined using two-sided Wilcoxon rank-sum tests.

Validations of the RBP-splicing regulation circuitry inferred by CASREL

We validated the RBP-AS regulatory circuits inferred by CASREL using RNA-seq data from HepG2 (hepatocellular carcinoma) and K562 (chronic myeloid leukemia) cells upon knock down of 262 RBPs from the ENCODE project [31]. Differential splicing events induced by RBP knockdown were compared with CASREL predictions across 9 datasets. For most RBPs, the regulatory target validation rates ranged from 45% to 75% across diverse contexts (Fig. 3A). Random permutations of the RBP-AS circuitry reduced validation rates to ~10% or lower (Fig. 3A), demonstrating the specificity of CASREL predictions. As an additional reference, we examined CLIP-seq datasets of 54 RBPs from POSTAR3 [32]. For many RBPs, only small proportions of their binding target RNAs exhibited differential splicing upon RBP knockdown in HepG2 or K562 cell (Fig. 3A), showing lower validation rates than CASREL predictions (Fig. 3A). These results suggest that CLIP-derived binding annotations do not necessarily indicate splicing regulation, as expected given the complexity of splicing regulation and off-targets inherent to CLIP-seq [33-35].

Notably, the validation datasets were from K562 and HepG2 cells, which differ from most cell types in the scRNA-seq datasets used for CASREL inference of AS regulatory circuits. In fact, cross validation between the AS events upon RBP knockdown in K562 and HepG2 yielded only 40-65% overlap (Fig. S4). Thus, given the cell type- and context-dependence of splicing regulation, the validation rates observed for CASREL predictions already provide strong support for its robustness in capturing context-specific AS regulators directly from scRNA-seq data.

Furthermore, we assessed CASREL predictions against RBP-RNA binding profiles from CLIP-seq of 97 RBPs [32], assuming that splicing regulation often requires direct binding. Supporting rates, defined as the proportion of CASREL-predicted targets with CLIP-detected binding of the RBPs, ranged from 10% to 60% across datasets (Fig. 3B), which dropped sharply upon random shuffling

of the RBP-AS circuitry (Fig. 3B). Nevertheless, by design, CASREL also captures RBP-AS regulation without direct RBP-RNA binding, potentially mediated via cofactors, thereby providing a more comprehensive view of the splicing regulatory landscape.

Finally, we compared CASREL with existing computational approaches, including SFpointer [36], Sparse Adjusted Motif-Adjusted Motif (SAM-AM) model [37], and DeepRBP [38]. Benchmarking across 9 datasets demonstrated that CASREL outperformed existing methods in both AS prediction accuracy and validation rates against RBP knockdown data (Fig. 3A, C). Notably, although DeepRBP accurately predicted AS profiles with bulk RNA-seq of tumor samples in TCGA (Fig. 3C), its predictions of RBP regulators for AS events were poorly supported by RBP knock-down experiments (< 25%), far below CASREL (Fig. 3A). Importantly, both SAM-AM and SFpointer rely on prior knowledge such as CLIP binding sites or sequence motifs and thus cover only a limited subset of RBPs. Together, these analyses highlight CASREL's unique capability to achieve both broad coverage and high-resolution discovery of *de novo*, cell type-specific splicing regulators.

Figure 3

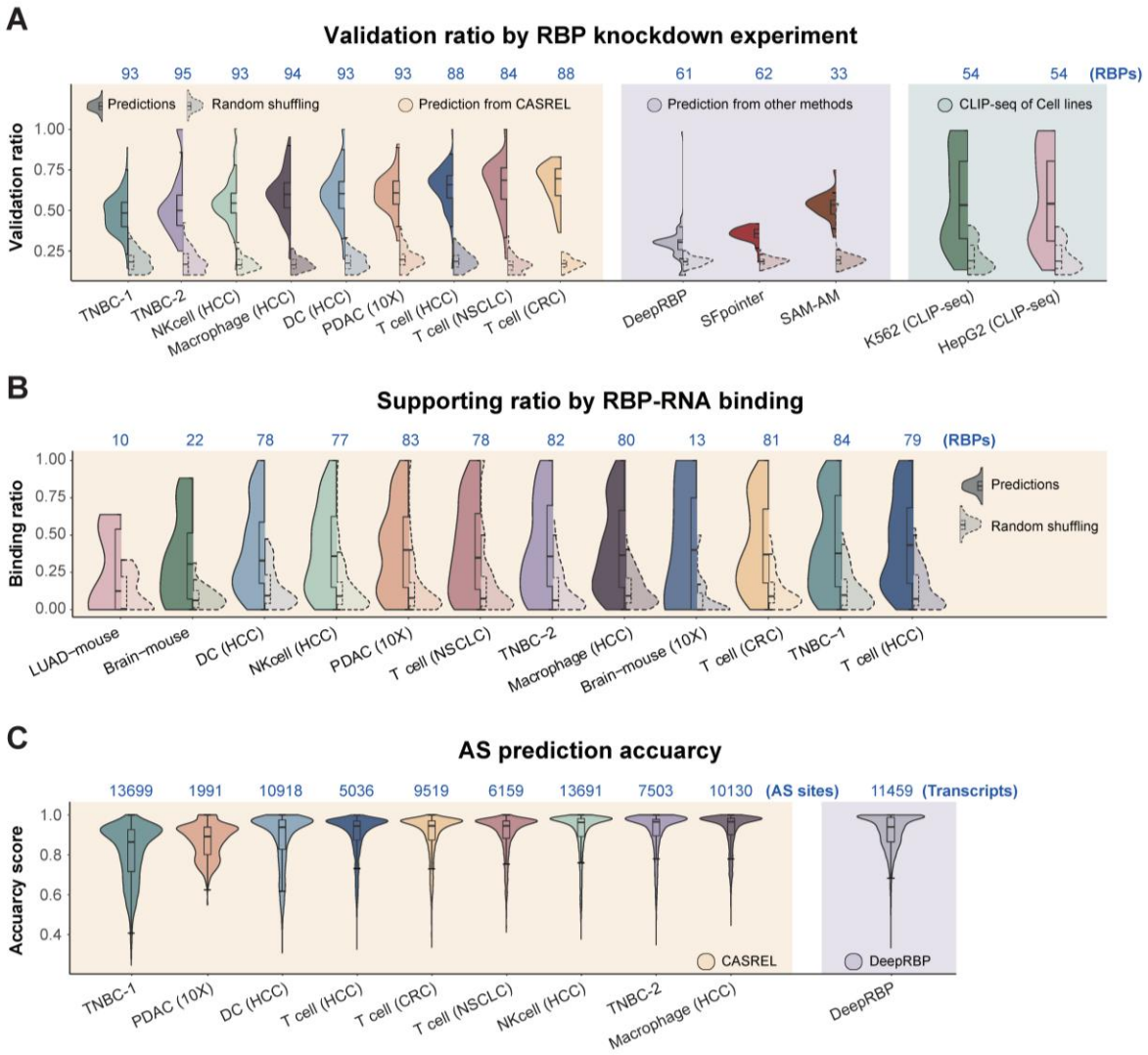


Figure 3. Experimental validations of the RBP-AS circuitry predicted by CASREL.

(A) Validation of CASREL predictions by RBP knockdown experiments. Violin plots and embedded box plots show the distributions of validation rates for CASREL predictions (9 datasets), existing computational methods, and RBP-RNA binding events identified by CLIP-seq data (2 cell lines). For each method, experimental validation (left) is compared with random controls (right; generated by randomly assigning the same number of AS events per RBP). For each data point in the violin plots, the y-axis indicates the proportion of predicted AS events for a given RBP that exhibit significant splicing changes ($|\Delta PSI| > 0.1$, FDR < 0.05) upon RBP knockdown in K562 or HepG2 cells. The number of RBPs subjected to knockdown experiments and included in each plot is indicated above.

(B) Percentages of RBP-AS circuits in 12 datasets supported by CLIP-seq experiments. For each dataset, CASREL predictions (left) are contrasted with a null distribution (right) generated by random reassignment of AS targets while preserving target counts for each RBP. The y-axis shows the fraction of AS events per RBP with CLIP-seq-validated binding. Numbers above indicate the number of RBPs with CLIP-seq data.

(C) Violin plots show splicing profile prediction accuracy on held-out test sets for CASREL (yellow, 9 datasets) and DeepRBP (purple). Accuracy of AS profile predictions were evaluated using consistent metrics and normalized across methods (see Methods). Numbers above each violin plot indicate the sample size, i.e., numbers of AS events.

Applications of CASREL in inferring cell-specific AS regulatory circuitry

To delineate cell-specific splicing regulatory programs, we applied CASREL on single-cell RNA-seq data from different tumor-infiltrating immune cells in HCC [39] and CRC [40]. As expected, the RBP-AS regulatory circuits inferred by CASREL displayed substantial heterogeneity across cell types (Fig. S5A), consistent with the well-recognized variability of splicing regulation across cell types, tissues, and environmental contexts [12, 41]. Notably, the same cell types (e.g., $CD4^+$ or $CD8^+$ T cells) from HCC and CRC tumors in independent studies showed greater similarity in their RBP-AS circuits than different cell types from the same tumor microenvironment (Fig. 4A, S5B), highlighting both the accuracy and low batch bias of CASREL.

We examined the splicing regulation of STAT3 transcripts in T cells as an example. As a key mediator of IL-6 signaling and regulator of cell proliferation and anti-tumor immunity [42], STAT3 has two major isoforms: the full-length oncogenic STAT3α and tumor-suppressive STAT3β generated by exon 23 skipping [43] (Fig. 4B). CASREL identified both conserved and cell type-specific regulators of STAT3β formation in HCC and CRC tumor-derived T cells, most of which have been reported in literature or supported by RBP knockdown experiments (Fig. 4C, D). Among the conserved regulators in both $CD4^+$ and $CD8^+$ T cells from HCC and CRC tumors, PCBP1 strongly promoted STAT3β formation, consistent with prior evidence that PCBP1 binds STAT3 exon 23 to enhance STAT3β generation in multiple cell lines [44] (Fig. 4C, D). GRSF1 and TRA2A acted as inhibitors of STAT3β, supported by RBP knockdown data in ENCODE (Fig. 4C, D, S6). Cell type-specific regulation patterns also emerged, showing promotive or inhibitory RBPs only in $CD4^+$ or $CD8^+$ T cells (Fig. 4C, D). For example, PCBP2, which shares similar RNA-binding preferences with PCBP1 [45], inhibited STAT3β formation selectively in $CD8^+$ T cells. Notably, the regulators of STAT3β in $CD4^+$ and $CD8^+$ T cells are highly consistent between HCC and CRC (Fig.

4E, F), supporting the cell type-dependency of RBP–AS regulation while indicating minimal batch bias in CASREL.

Figure 4

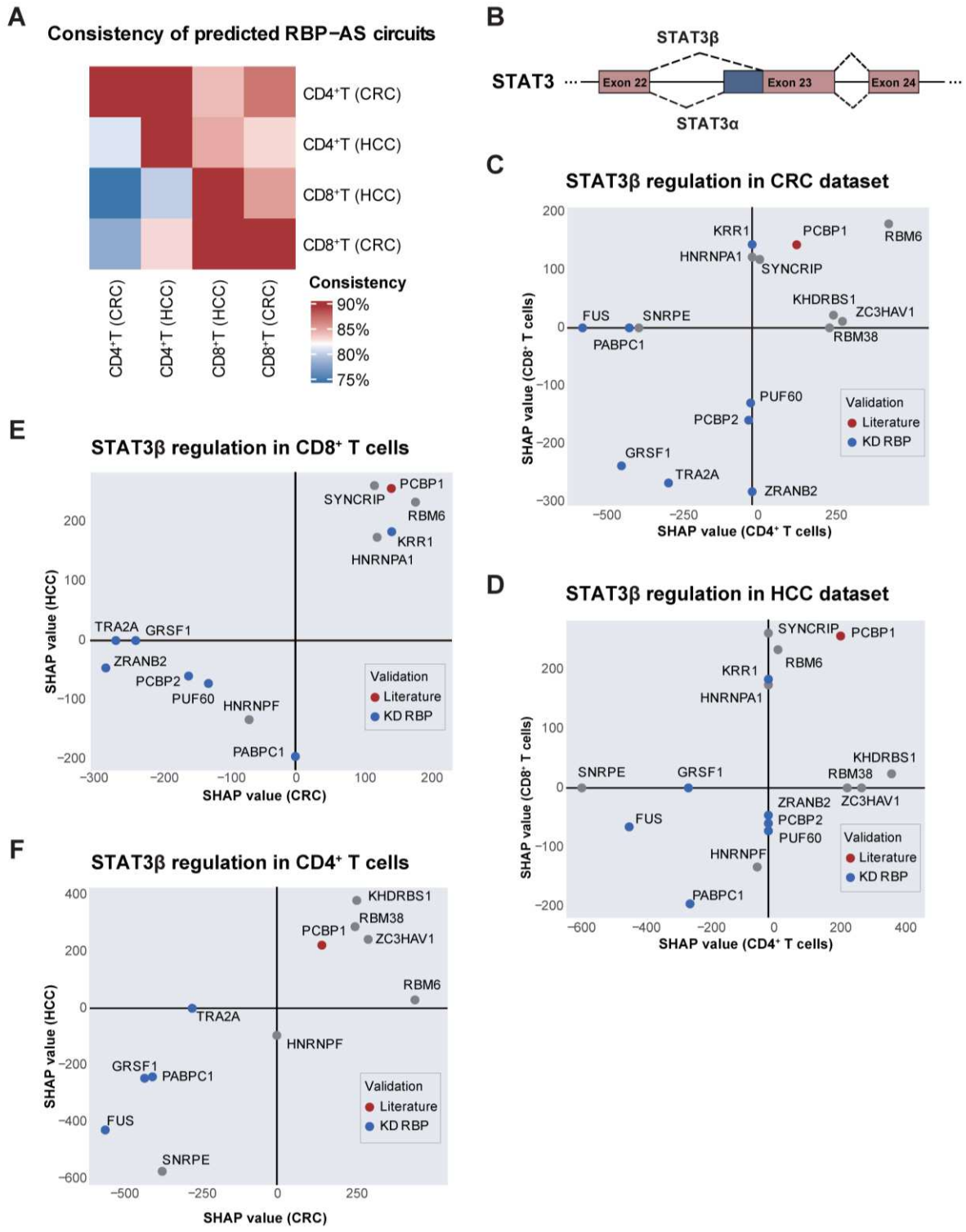


Figure 4. Cell type-specific RBP-splicing regulatory circuits in tumor microenvironments.

(A) Heatmap showing consistency of CASREL-predicted RBP-AS circuit between $CD8^+$ and $CD4^+$ T cells across different tumor microenvironment datasets. Color intensity reflects the proportion of shared circuits normalized to the reference cell type (horizontal axis).

(B) Schematic of alternative splicing at STAT3 exon 23, generating STAT3α and STAT3β isoforms.

(C, D) SHAP-derived regulatory effects of RBPs in CD4+ versus CD8+ T cells from CRC (**C**) and HCC (**D**) tumor microenvironments. Positive SHAP values denote promotion of STAT3β inclusion, and negative values indicate inhibition. RBPs are color-coded by validation: red, literature-validated STAT3β regulators; blue, RBPs functionally validated via knockdown in HepG2/K562 cells.

(E, F) Scatter plots illustrating the distribution of SHAP values for predicted STAT3β regulators in (**E**) $CD8^+$ and (**F**) $CD4^+$ T cells across CRC and HCC contexts. Axes represent SHAP values derived from HCC (x-axis) and CRC (y-axis) analyses. RBP coloring reflects validation status as in (**C, D**).

Discussion

Cell type-specific AS programs are major drivers of splicing heterogeneity across tissues, developmental stages, and microenvironmental conditions. Deciphering these programs is therefore essential for understanding the post-transcriptional machineries underlying key biological processes. However, most existing computational approaches rely on CLIP-derived binding annotations or predefined RBP binding motifs, thereby being restricted to a few preselected RBPs in limited model cell lines. Moreover, bulk RNA-seq averages across heterogeneous populations, diluting signals from both RBP expression and AS events [46] and masking cell type-dependent RBP-AS regulatory machineries [47]. To overcome these limitations, CASREL leverages the unique properties of scRNA-seq for *de novo*, cell type-specific inference of RBP-AS regulation circuits without prior knowledge of RBP functions, RNA-binding motifs, or CLIP annotations.

Single-cell transcriptomics, particularly full length-coverage sequencing platforms like Smart-seq, enables joint profiling of both gene expression and AS landscapes at single-cell resolution. Unlike bulk RNA-seq, where AS signals are averaged across heterogeneous populations, single-cell splicing profiles often display polarized patterns, with most genes dominated by a single transcript isoform [17]. This may reflect temporal coupling between transcription bursts and RNA splicing. During short periods of transcriptional bursts, spliceosome assembly likely occurs within narrow time windows that may allow a single dominating regulatory machinery [48]. This model

is supported by live-cell imaging, which demonstrates synchronized splicing kinetics during transcriptional bursts [49], and explains the reduced splicing complexity captured in single-cells compared to population-level averages. Moreover, transcriptional bursting itself may serve as a tunable mechanism for cell type-specific splicing, in which burst kinetics could be reprogramed in specialized cells to favor slower-acting regulatory complexes on weak cis-elements [50].

The unique features of single-cell RNA-seq data confer two key advantages for splicing analysis: (i) cleaner AS profiles presumably capturing dominant regulatory schemes, and (ii) direct access to cell type-specific splicing landscapes. Designed to leverage these unique advantages, CASREL integrates single-cell splicing profiles with RBP expression data, using machine learning to predict AS states and identify context-specific RBP-AS associations. This framework applies a novel four-state discretization strategy, which (i) accommodates the strong polarization of single-cell isoform usage, (ii) mitigates the inherent sparsity of single-cell data with from high dropout rates, and (iii) preserves potential biological information encoded in the pattern of data dropouts [2, 20].

Multiple lines of evidence support the robustness of CASREL. It performs consistently across diverse cell types and tissues, with predictions validated by RBP knockdown experiments and CLIP-seq data. Importantly, CASREL reveals pronounced context-dependent variation in splicing regulation, underscoring the critical role of cellular states and environment factors in shaping splicing landscapes. Such AS regulatory complexity is largely obscured in bulk RNA-seq analyses, highlighting the unique power of single-cell resolution for dissecting AS regulatory circuitry.

In conclusion, CASREL provides a powerful framework for reconstructing splicing regulatory circuitry at single-cell resolution, offering a comprehensive view of context-dependent splicing regulation. Further extensions could incorporate additional regulatory factors of RNA splicing, such as non-coding RNAs, chromatin accessibility, or post-translational modifications of RBPs. Coupled with emerging technologies like spatial transcriptomics, CASREL has the potential to map splicing regulation within tissue architecture, offering deeper insights into the interplay between splicing programs and cellular phenotypes.

Methods

Data Source

Nine sets of single-cell RNA-seq data generated by the SMART-seq platform and 10X platform were downloaded from the NCBI SRA Database (Table S1). These datasets were previously published for studies of TNBC, NSCLC, tumor-associated immune cells, mouse embryonic liver and brain. The raw reads were mapped to the reference genomes (human hg19 or hg38, according to the original studies, or mouse mm10) with hisat2, generating BAM files. SCASL was then used to obtain the AS profiles of these samples from the BAM files.

In addition, we obtained RBP binding site information from the publicly accessible database POSTAR3[32]. POSTAR3 encompasses an extensive collection of 1499 CLIP-seq datasets, representing 348 RBPs. These datasets span across 10 distinct CLIP-seq technologies, namely HITS-CLIP, PAR-CLIP, iCLIP, eCLIP, iCLAP, urea-iCLIP, 4sU-iCLIP, BrdU-CLIP, Fr-iCLIP, and PIP-seq. The data covers seven species, including human, mouse, zebrafish, fly, worm, Arabidopsis, and yeast. Notably, POSTAR3 is the most comprehensive and up-to-date database that consolidates RBP binding site information from various CLIP-seq technologies and multiple species.

The RNA-seq data utilized in this study, obtained from the ENCODE database, consists of transcriptome sequencing following the independent knockdown of 356 RBPs in two human cell lines: K562 (chronic myeloid leukemia cell line) and HepG2 (hepatocellular carcinoma cell line). To align the raw sequencing data and generate a BAM file, we employed hisat2 and hg38 as the reference genome.

AS profiles are generated from BAM files using SCASL, allowing us to obtain the AS probabilities of all samples. Additionally, we utilized featureCounts to extract the expression information from the BAM file, enabling us to generate the gene expression profile for each sample.

The CASREL pipeline

Data preprocessing

Discretization of AS profile: In single-cell AS profile, AS probabilities represented as relative abundance values ranging from 0 to 1, with a majority of data exhibiting a polarized distribution. For each splicing event, the AS probability value was converted into one of the following

classifications: non-spliced (splicing probability < 0.4), neutral (0.4-0.6), spliced (> 0.6), and NA (missing coverage).

k-fold partition of datasets: To enhance the reliability and stability of RBP-AS regulatory pairs, we utilized a five-fold cross-validation (k=5) approach, dividing all samples in the dataset into five equal parts. In each iteration, one part was selected as the testing set while the remaining four parts served as the training set. This ensured that each training and testing set combination was distinct. This method is widely applied in computer science and machine learning to evaluate model performance and mitigate the risk of overfitting. For each AS event, we trained a model on each training set group, resulting in five different predictive models.

Modeling for AS event

GBDT: We use GBDT (Gradient Boosting Decision Trees) as a prediction model for categorical AS profiles from RBP expressions. The training iteration process of gradient boosting can be described as the following greedy optimization algorithm:

1) Learning the simplest constant model:

$$F_0(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \quad (1)$$

2) Iterating for *m* times: the following process is performed to learn a series of weak learners, denoted as h_m, such that their sum forms an ideal ensemble model. In gradient boosting algorithm, weak learners or base learners are small models used to fit the model prediction residuals in each iteration. The weak learners employed by GBDT are decision tree models. For the m-th step, the optimization objective is to minimize the residual as follows:

$$F_m(x) = F_{m-1}(x) - \gamma \sum_{i=1}^{n} \nabla_{F_{m-1}} L\big(y_i, F_{m-1}(x_i)\big) \quad (2)$$

This is an intractable functional optimization problem that cannot be directly solved. Therefore, GBDT tackles this problem using a two-step coordinate gradient descent approach. Specifically, the loss function is treated as a function of the ensemble learning model obtained in the previous step, $F_{m-1}(x)$, and pseudo-residuals are computed by calculating the gradients. According to the definition of gradients, they represent the specific direction and relative magnitude of the residual for each sample:

$$r_{im} = -\left[\frac{\partial L\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \text{for } i = 1, \cdots, n \tag{3}$$

Next, we employ the conventional gradient descent algorithm to solve for a weak learner $h_m(x)$ that fits the computed pseudo-residuals. Since the pseudo-residuals represent the direction and relative magnitude of the residuals, we need an appropriate regularization multiplier γ_m to scale them. It is worth noting that this multiplier can be obtained by solving a one-dimensional optimization problem

$$\gamma_m = \text{argmin}_\gamma \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right) \tag{4}$$

Next, we integrate the newly obtained weak learner into the model by weighted summation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{5}$$

Light GBM: In our study, we specifically employed LightGBM as the implementation of GBDT. LightGBM incorporates improvements and optimizations over the basic GBDT framework, aiming to enhance training and prediction efficiency while reducing memory usage [21]. Specifically, LightGBM possesses the following characteristics:

1) Histogram-based binning strategy: LightGBM adopts a histogram-based decision tree algorithm that discretizes continuous features into bins, introducing regularization through bin quantity control while reducing feature dimensionality, significantly decreasing computational and memory overhead. To optimize training efficiency while preserving information gain fidelity, LightGBM implement Gradient-based One-Side Sampling (GOSS). This technique leverages the established principle that samples with larger gradient magnitudes exert disproportionately high influence on information gain calculations. To address the computational challenges of high-dimensional sparse feature spaces, LightGBM implement Exclusive Feature Bundling (EFB). This technique exploits the observation that in many sparse datasets, multiple features exhibit mutual exclusivity (i.e., they rarely take non-zero values simultaneously). EFB reduces dimensionality by strategically bundling such exclusive features into composite features through a two-phase process: First, a greedy approximation algorithm identifies near-optimal bundles by iteratively grouping features with minimal conflict counts, circumventing the NP-hard nature of the optimal bundling problem. Subsequently, a value reassignment procedure merges features within each bundle by offsetting their original value ranges. For example, features with ranges [0, 10] and [0,

20] are merged into a new feature with range [0, 30], where original values of the second feature are mapped to [10, 30]. This offset-based merging guarantees zero information loss while eliminating histogram computation conflicts during gradient boosting operations.

2) Leaf-wise growth strategy: In contrast to traditional level-wise growth strategies, LightGBM employs a leaf-wise growth strategy. It selects nodes that minimize the loss the most for splitting, thereby enhancing training efficiency.

3) Support for parallelization: LightGBM efficiently handles large-scale data and supports parallel processing, resulting in improved training speed.

The features above enable us to rapidly train a large number of high-performing and interpretable GBDT models. Due to the similarity between decision trees and human decision-making processes, decision trees offer strong interpretability. However, if decision trees become overly complex (e.g., deep trees with numerous splitting features), they are prone to overfitting. Gradient boosting, as an ensemble learning algorithm, significantly reduces the degree of overfitting. In gradient boosting, to keep the training cost of weak learners low and ensure model simplicity while reducing the risk of overfitting, trees are typically limited to 4 to 8 levels. This restriction is achieved by penalizing tree depth through a regularization loss term. Additionally, other regularization terms can be employed to limit the complexity of each tree from different aspects, such as penalizing excessive numbers of leaf nodes.

Model interpretation

SHAP value: Based on game theory, SHAP (SHapley Additive exPlanations) assigns a value to each feature to represent its contribution to the model output [26]. Its principle is to compute the contribution of a feature to the model output when it is added as a predictor. By employing a greedy algorithm that progressively adds features, SHAP values can be represented as parameters of a linear explanation model:

$$g(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \, x_i \tag{6}$$

In classical game theory, SHAP values can be computed as follows:

$$\phi_i(f) = \sum_{S \subseteq N - \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{7}$$

where N represents the set of features, S is any subset that does not include feature i, $f(S)$ represents the model's output when using feature set S (taking the expectation for the excluded feature), and $f(S \cup \{i\})$ represents the model's output when including feature i.

Calculation of SHAP values for tree ensemble models: For tree ensemble models, the calculation of SHAP values involves traversing the nodes of the decision trees and considering the contribution of each node to the model's output. The SHAP value calculation method mentioned in reference [51] employs a weighted path algorithm, which fundamentally involves summing the weighted contributions of each feature along each decision path to obtain the SHAP values for each feature.

$$\phi_i(f) = \sum_{p \in \prod(i)} \frac{1}{|\prod(i)|} \cdot \Delta_{f,p} \tag{8}$$

In this context, $\phi_i(f)$ represents the SHAP value of feature f for sample I. $\Pi(i)$ denotes the set of all decision paths that include sample i, where the cardinality of $\Pi(i)$ represents the cumulative weight of each path. $\Delta_{f,p}$ signifies the variation of feature f along path p.

Aggregation of SHAP Values: After obtaining the per-sample, per-class, and per-feature SHAP values, we performed an aggregation operation to obtain accumulated SHAP values across all samples and classes within the same dataset. For sample-level aggregation, a simple averaging operation is employed. For class-level aggregation, we use the optimal voting method, where the aggregated SHAP value is determined by the SHAP value corresponding to the predicted class of the final model. By performing these aggregation operations, we obtain the SHAP values for a single RBP-AS pair. These values serve as crucial quantitative measures for conducting AS regulatory analysis.

Filtering of results and k-fold validation: For each of the k-fold training process, we conducted a rigorous screening to identify qualified models based on a prediction accuracy > 0.85. We further filtered the models by ranking the top 20 RBPs according to their contribution, as determined by the SHAP value, in the corresponding event. Additionally, we retained only those RBPs with a SHAP value exceeding 10. By applying these criteria, we obtained RBP-AS regulatory pairs for each fold.

We then obtained the intersection of RBP-AS regulatory pairs across all five folds, while ensuring consistent regulatory directions. Moreover, we calculated the final SHAP value for each

consensus regulatory pair by taking the summed average of the SHAP values across the five folds. This final SHAP value served as a metric to assess the significance of the RBP's impact on the AS event.

Benchmarking Performances of CASREL

Performance comparison of different machine learning models

Given the multi-class nature of our prediction task, we departed from traditional binary classification metrics such as ROC curves and AUC values, instead adopting accuracy scores as our primary evaluation metric. In each model, $\widehat{y_i}$ is the predicted value of the i-th sample, $\widehat{y_i}$ is the true value in the test set corresponding to the sample, and the prediction accuracy score in $n_{samples}$ is:

$$accuarcy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\widehat{y_i} = y_i) \quad (9)$$

To benchmark the performance of diverse predictive models from different methods on a comparable scale, we employed the canonical evaluation metrics from each model's original publication and subsequently normalized these scores to a common [0,1] interval. This approach allowed for an equitable comparison of predictive accuracy despite fundamental differences in the models' tasks and outputs:

By comparing the prediction accuracy across these three models, we gained insights into their respective performance. For feature importance interpretation, we employed distinct approaches tailored to each model's characteristics. Specifically, SHAP values were calculated for the non-linear GBDT and FCN models to quantify individual feature contributions to predictions. For the linear LR model, the absolute LR coefficients were analyzed to assess global feature importance, where the coefficient magnitude reflects the feature's impact on the log-odds of class prediction. In addition, we validated the plausibility of the interpretations of each model by calculating the percentage of RBPs in different models that exhibited monotonic regulatory effects across all AS events in multiple datasets (an RBP was classified as monotonic if > 90% of its regulatory effects exhibited consistent promotion or inhibition).

Assessment of CASREL performance stability

The 5-fold cross-validation strategy yielded five independent model instances, each with different training and test set compositions. Our stability assessment focused on two critical dimensions: prediction performance stability and interpretability stability:

1) Stability of model prediction performance. By calculating the prediction accuracy and distribution characteristics of each model instance and the coefficient of variation (CV), we assessed the accuracy consistency of the model across different training / test set partitions. The prediction accuracy obtained for each of the K-folds is x_i, and the CV is:

$$CV = \frac{\sqrt{\frac{1}{4}\sum_{i=1}^{5}\left(x_i - \frac{1}{5}\sum_{i=1}^{5} x_i\right)^2}}{\frac{1}{5}\sum_{i=1}^{5} x_i} \times 100\% \quad (10)$$

2) Stability of model interpretability. SHAP values are used to interpret the model. SHAP values and their distribution are systematically analyzed for each model instance to verify the consistency and reliability of the model's interpretability across different data subsets. We aligned each splicing event and predicted regulatory factor within each k-fold and calculated the Pearson correlation coefficient between the predicted results and the contribution (SHAP value) of each RBP-AS regulatory circuitry within each fold.

Validations of CASREL predictions by RBP functional experiments

To validate the regulatory interactions predicted by CASREL, we implemented a comprehensive experimental validation framework incorporating both transcriptome-wide and molecular-level analyses. We obtained transcriptome profiles following RBP knockdown experiments in two cell lines (K562 and HepG2) from the ENCODE database. Differential splicing patterns were systematically analyzed using SCASL methodology, followed by statistical assessment through Wilcoxon rank sum tests. The predictive accuracy of CASREL was quantified by calculating the reproducibility of the inferred regulatory events in these perturbation experiments, defined as the proportion of correctly predicted events among significantly differentially regulated splicing events.

To establish a genome-wide map of RBP-RNA interactions, we integrated CLIP-seq data from the POSTAR3 database. We performed spatial colocalization analysis to evaluate the physical proximity between predicted RBP binding sites and their putative AS targets. This analysis

quantified the binding probability within the CASREL-predicted regulatory network by calculating the proportion of CLIP-seq-validated binding events among the predicted RBP-AS regulatory pairs. To account for technical and biological variability, we assessed the consistency of RBP binding patterns across different cell lines and sequencing platforms by computing the concordance rate of binding events between distinct experimental conditions.

We further benchmarked CASREL against existing computational approaches for splicing regulator prediction, encompassing three major algorithmic paradigms: CLIP-seq-based methods focusing on physical interactions (exemplified by SFPointer), sequence conservation-based approaches (such as SAM-AM), and machine learning models trained on population-scale sequencing data (like DeepRBP). Using standardized datasets and analytical pipelines as described in the respective publications, we conducted comparative analyses of prediction accuracy and experimental validation rates across all methods. Performance metrics were evaluated following established protocols, with the concordance between computational predictions and experimental RBP binding data.

Differential alternative splicing analysis

We conducted differential splicing analysis on the filtered raw AS profiles using the Wilcoxon rank sum test. To ensure robustness and avoid potential imputation errors, we utilized the unimputed AS ratio matrix for calculating differential splicing events. In the case of bulk RNA-seq datasets involving RBP knockdown experiments, we compared the differences in AS ratios between the RBP knockdown samples and corresponding non-knockdown control samples within the same cell line. We considered events with adjusted p-values below 0.05 as differentially spliced events. Finally, we mapped each identified event back to the gene with the highest inclusion and the closest proximity based on the positional relationship between the junction and the gene on the chromosome.

Data Availability

All single-cell data utilized in this manuscript were sourced from publicly available databases, which was deposited in the Gene Expression Omnibus (GEO). The accession numbers associated with the datasets are as follows: GSE123837, GSE118390, GSE152607, GSE71585, GSE205013, GSE95315, GSE98638, GSE108989, and GSE99254.

Code Availability

Codes for all the data analyses in this study are available in github (https://github.com/xryanglab/CASREL).

Acknowledgements

The authors would like to thank the supports from the Tsinghua University Branch of China National Center for Protein Sciences (Beijing) and Tsinghua University Technology Center for Protein Research, including the core facilities of Biocomputing, Genome Sequencing and Analysis at Tsinghua University. This work was funded by the National Key Research and Development Program of China (2025YFC3409300, 2023YFC3043300), the National Natural Science Foundation of China (32330022 and W2411012), the Natural Science Foundation of Chongqing (CSTB2025NSCQ-GPX1078), the Special Support Program for Postdoctoral Research Projects of Chongqing (2024CQBSHTB2006) and the Medical Research Innovation Foundation of Chongqing General Hospital (Y2024YXYJZDXM02).

Ethical approval

Ethical approval is not needed for this study.

Declaration of Interests

The authors declare no competing interests.

References

1. Marasco, L.E. and A.R. Kornblihtt, *The physiology of alternative splicing.* Nature Reviews Molecular Cell Biology, 2023. **24**(4): p. 242-254.
2. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nature genetics, 2008. **40**(12): p. 1413-1415.
3. Lam, S.D., et al., *Biological impact of mutually exclusive exon switching.* Plos Computational Biology, 2021. **17**(3): p. e1008708.

4. Martín, G., et al., *Alternative splicing landscapes in Arabidopsis thaliana across tissues and stress conditions highlight major functional differences with animals.* Genome biology, 2021. **22**(1): p. 1-26.

5. Marquez, Y., et al., *Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity.* Genome research, 2015. **25**(7): p. 995-1007.

6. Bush, S.J., et al., *Alternative splicing and the evolution of phenotypic novelty.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2017. **372**(1713): p. 20150474.

7. Kastner, B., et al., *Structural insights into nuclear pre-mRNA splicing in higher eukaryotes.* Cold Spring Harbor perspectives in biology, 2019. **11**(11): p. a032417.

8. Wu, S., et al., *ASCancer Atlas: a comprehensive knowledgebase of alternative splicing in human cancers.* Nucleic Acids Research, 2023. **51**(D1): p. D1196-D1204.

9. Lovci, M.T., et al., *Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges.* Nature structural & molecular biology, 2013. **20**(12): p. 1434-1442.

10. Wright, C.J., C.W. Smith, and C.D. Jiggins, *Alternative splicing as a source of phenotypic diversity.* Nature Reviews Genetics, 2022. **23**(11): p. 697-710.

11. Witten, J.T. and J. Ule, *Understanding splicing regulation through RNA splicing maps.* Trends in Genetics, 2011. **27**(3): p. 89-97.

12. Fu, X.-D. and M. Ares Jr, *Context-dependent control of alternative splicing by RNA-binding proteins.* Nature Reviews Genetics, 2014. **15**(10): p. 689-701.

13. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-476.

14. Raj, B. and B.J. Blencowe, *Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles.* Neuron, 2015. **87**(1): p. 14-27.

15. König, J., et al., *Protein–RNA interactions: new genomic technologies and perspectives.* Nature Reviews Genetics, 2012. **13**(2): p. 77-83.

16. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nature reviews Molecular cell biology, 2017. **18**(7): p. 437-451.

17. Xiang, X., et al., *Interrogations of single-cell RNA splicing landscapes with SCASL define new cell identities with physiological relevance.* Nature Communications, 2024. **15**(1): p. 2164.

18. Zhang, X., et al., *Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex.* Cell, 2016. **166**(5): p. 1147-1162. e15.

19. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nat Rev Mol Cell Biol, 2017. **18**(7): p. 437-451.

20. Gulati, G.S., et al., *Single-cell transcriptional diversity is a hallmark of developmental potential.* Science, 2020. **367**(6476): p. 405-411.

21. Ke, G., et al., *Lightgbm: A highly efficient gradient boosting decision tree.* Advances in neural information processing systems, 2017. **30**.

22. Li, J., et al., *Tnt: An interpretable tree-network-tree learning framework using knowledge distillation.* Entropy, 2020. **22**(11): p. 1203.

23. Hancock, J. and T.M. Khoshgoftaar. *Leveraging lightgbm for categorical big data*. in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*. 2021. IEEE.

24. Zhang, C., et al., *An up-to-date comparison of state-of-the-art classification algorithms.* Expert Systems with Applications, 2017. **82**: p. 128-150.

25. Christoph, M., *Interpretable machine learning: A guide for making black box models explainable.* 2020.

26. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 2017. **30**.

27. Prochazka, L., R. Tesarik, and J. Turanek, *Regulation of alternative splicing of CD44 in cancer.* Cellular signalling, 2014. **26**(10): p. 2234-2239.

28. Yae, T., et al., *Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell.* Nature communications, 2012. **3**(1): p. 883.

29. Loh, T.J., et al., *hnRNP L inhibits CD44 V10 exon splicing through interacting with its upstream intron.* Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2015. **1849**(6): p. 743-750.

30. Xu, Y., et al., *Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing.* Genes & development, 2014. **28**(11): p. 1191-1203.

31. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57.

32. Zhao, W., et al., *POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins.* Nucleic acids research, 2022. **50**(D1): p. D287-D294.

33. Jackson, A.L., et al., *Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity.* Rna, 2006. **12**(7): p. 1179-1187.

34. Maida, Y., et al., *Off-target effect of endogenous siRNA derived from RMRP in human cells.* International Journal of Molecular Sciences, 2013. **14**(5): p. 9305-9318.

35. Jackson, A.L., et al., *Expression profiling reveals off-target gene regulation by RNAi.* Nature biotechnology, 2003. **21**(6): p. 635-637.

36. Lobato-Fernandez, C., et al., *A Systematic Identification of RNA-Binding Proteins (RBPs) Driving Aberrant Splicing in Cancer.* Biomedicines, 2024. **12**(11): p. 2592.

37. Gupta, K., et al., *Improved modeling of RNA-binding protein motifs in an interpretable neural model of RNA splicing.* Genome Biology, 2024. **25**(1): p. 23.

38. Sancho, J., et al., *DeepRBP: A novel deep neural network for inferring splicing regulation.* bioRxiv, 2024: p. 2024.04. 11.589004.

39. Zheng, C., et al., *Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing.* Cell, 2017. **169**(7): p. 1342-1356. e16.

40. Zhang, L., et al., *Lineage tracking reveals dynamic relationships of T cells in colorectal cancer.* Nature, 2018. **564**(7735): p. 268-272.

41. Keppetipola, N., et al., *Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2.* Critical reviews in biochemistry and molecular biology, 2012. **47**(4): p. 360-378.

42. Yu, H., D. Pardoll, and R. Jove, *STATs in cancer inflammation and immunity: a leading role for STAT3.* Nature reviews cancer, 2009. **9**(11): p. 798-809.

43. Caldenhoven, E., et al., *STAT3β, a splice variant of transcription factor STAT3, is a dominant negative regulator of transcription.* Journal of Biological Chemistry, 1996. **271**(22): p. 13221-13227.

44. Wang, X., et al., *PCBP1 inhibits the expression of oncogenic STAT3 isoform by targeting alternative splicing of STAT3 exon 23.* International journal of biological sciences, 2019. **15**(6): p. 1177.

45. Ishii, T., et al., *PCBP1 and PCBP2 both bind heavily oxidized RNA but cause opposing outcomes, suppressing or increasing apoptosis under oxidative conditions.* Journal of Biological Chemistry, 2020. **295**(34): p. 12247-12261.

46. Kharchenko, P.V., *The triumphs and limitations of computational methods for scRNA-seq.* Nature methods, 2021. **18**(7): p. 723-732.

47. He, S., et al., *The nexus between RNA-binding proteins and their effectors.* Nature Reviews Genetics, 2023. **24**(5): p. 276-294.

48. Fingerhut, J.M., et al., *Co-transcriptional splicing facilitates transcription of gigantic genes.* PLoS genetics, 2024. **20**(6): p. e1011241.

49. Drexler, H.L., K. Choquet, and L.S. Churchman, *Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores.* Molecular cell, 2020. **77**(5): p. 985-998. e8.

50. Meeussen, J.V. and T.L. Lenstra, *Time will tell: comparing timescales to gain insight into transcriptional bursting.* Trends in Genetics, 2024. **40**(2): p. 160-174.

51. Lundberg, S.M., G.G. Erion, and S.-I. Lee, *Consistent individualized feature attribution for tree ensembles.* arXiv preprint arXiv:1802.03888, 2018.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CASRELSupplFigv3.3.docx
- SupplTable.docx