BMC Bioinformatics

RESEARCH

Open Access

Impact of U2-type introns on splice site prediction in *A. thaliana* species using deep learning

Espoir Kabanga[1,2*], Seonil Jee[1], Soeun Yun[1], Stephen Depuydt[3], Arnout Van Messem[4] and Wesley De Neve[1,2]

*Correspondence:
Espoir Kabanga
espoir.kabanga@ghent.ac.kr
[1]Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon 21985, Republic of Korea
[2]IDLab, Department of Electronics and Information Systems, Ghent University, Ghent 9000, Belgium
[3]Department of Health Care, HOGENT University of Applied Sciences and Arts, Ghent 9000, Belgium
[4]Department of Mathematics, University of Liège, Liège 4000, Belgium

Abstract

Background Splice site prediction in plant genomes poses substantial challenges that can be addressed using deep learning models. U2-type introns are especially useful for such studies given their ubiquity in plant genomes and the availability of rich datasets. We formulated two hypotheses: one proposing that short introns may enhance prediction effectiveness due to reduced spatial complexity, and another suggesting that sequences with multiple introns provide a richer context for splicing events.

Results Our findings demonstrate that (1) models trained on datasets containing shorter introns achieve improved effectiveness for acceptor splice sites, but not for donor splice sites, indicating a more nuanced relationship between intron length and splice site prediction than initially hypothesized, and (2) models trained on datasets with multiple introns per sequence show higher effectiveness compared to those trained on datasets with a single intron per sequence. Notably, among the 402 bp sequences analyzed, 72% contained single introns while 28% contained multiple introns for donor sites (36,399 versus 13,987 sequences), with similar proportions observed for acceptor sites (37,236 versus 14,112 sequences). These computational insights align with biological observations, particularly regarding the conserved spatial relationship between branch points and acceptor splice sites, as well as the synergistic effects of multiple introns on splicing efficiency.

Conclusions The obtained results contribute to a deeper understanding of how intronic features influence splice site prediction and suggest that future prediction models should consider factors such as intron length, multiplicity, and the spatial arrangement of splice-related signals.

Keywords *Arabidopsis thaliana*, CNN, Splice site prediction, U2-type introns

Background

In eukaryotic organisms, during messenger RNA (mRNA) processing, introns are excised, and exons are joined to form mature mRNA, which serves as a template for protein synthesis [10]. This process is known as splicing, and is carried out by the

spliceosome, a dynamic complex composed of small nuclear ribonucleoproteins and protein molecules.

The spliceosome recognizes specific sequences at the boundaries between introns and exons, known as splice sites, and removes the introns while ligating the exons together. Splice sites include both donor and acceptor sites. A donor splice site, located at the 5' end of an intron, is characterized by the dinucleotide `GT`, signaling the beginning of the intron to be removed during RNA splicing. An acceptor splice site, found at the 3' end of an intron, is marked by the dinucleotide `AG`, indicating the end of the intron to be excised. These sites play crucial roles in the accurate removal of introns and the subsequent joining of exons, facilitating the production of mature messenger RNA [13].

Accurate prediction of splice sites is a crucial element in gene expression analysis [35]. Introns can affect gene expression in plants and many other eukaryotes in a variety of ways [42]. U12-type and U2-type introns are two distinct classes of introns found in eukaryotic genomes [46]. They differ in terms of their spliceosomal machinery and splicing mechanisms [7].

U12-type introns are a less common class of introns, constituting a small fraction of introns in most eukaryotic genomes. They are spliced out by the minor spliceosome, a smaller and less well-understood spliceosome complex compared to the major spliceosome, which splices out U2-type introns. U12-type introns have distinct consensus sequences at the donor splice site (`AT-AC`, but many U12-type introns can also be `GT-AG`) and the branch point sequence.

U2-type introns, defined in this study by `GT-AG` boundaries at their splice sites [36], are the most prevalent type of introns in eukaryotic genomes, comprising over 99% of all introns in most eukaryotic species [46]. They are spliced out through the major spliceosome, a larger and more common spliceosome complex. The major spliceosome recognizes a highly conserved intron-exon junction sequence at the donor splice site, as well as a branch point sequence near the acceptor splice site [16].

Our focus on U2-type introns in the context of splice site prediction in *Arabidopsis thaliana* is motivated by several factors. First, the ubiquity of U2-type introns in plant genomes, including *Arabidopsis thaliana*, makes them an important topic for understanding the global splicing landscape and its regulatory mechanisms. Second, the vast amount of data available for U2-type introns facilitates the application of deep learning techniques, which require large datasets for effective model training. Third, the choice of *Arabidopsis thaliana* over other species is particularly strategic due to its well-annotated genome and its status as a model organism for dicots in plant genetics and molecular biology research [11].

Deep learning has emerged as an important tool in computational biology, revolutionizing various aspects of bioinformatics [5], including the prediction of splice sites. The complex nature of gene splicing, coupled with the vast amount of genomic data available, necessitates advanced computational techniques for the accurate prediction of splice sites. Deep learning models, for instance making use of convolutional neural networks (CNNs) [1–3, 53, 57], have shown remarkable proficiency in extracting complex features from DNA and RNA sequences, enabling them to discern subtle patterns crucial for splice site identification. However, U2-type introns, the predominant class of introns in eukaryotic genomes, pose unique challenges to splice site prediction [45] due to factors such as sequence variability [54], the presence of multiple potential splice sites,

and the complex regulatory elements associated with them [39]. Experimental evidence, such as the results presented in [28] on the ERECTA gene in *Arabidopsis thaliana*, further underscores the beneficial impact of multiple introns on gene expression. This study demonstrates that, while no single intron may be crucial for expression, the cumulative presence of multiple introns can substantially enhance mRNA accumulation and gene expression in an additive way. Incorporating these distinctive characteristics into deep learning models increases their precision and generalizability when dealing with U2-type intron-related splice site recognition. By systematically analyzing how intron features affect splice site predictability across thousands of sequences, computational approaches can complement experimental studies in identifying which sequence characteristics are associated with robust splicing signals [34, 40], ultimately facilitating more accurate splice site prediction for genome annotation and gene expression analysis.

Materials and methods

Dataset

The dataset used in this study was obtained from DRANetSplicer [33], a recent study on splice site prediction. This dataset, derived from the *Arabidopsis thaliana* genome, provides high-quality annotations of both donor and acceptor splice site sequences. Each sequence in the dataset is 402 bp long, consisting of 200 bp upstream and 200 bp downstream regions flanking the canonical splice site, defined by the dinucleotide `GT` for donor sites and `AG` for acceptor sites. In other words, the splice site is located precisely at the center of the sequence, specifically at positions 201 and 202. As described in DRANetSplicer, the datasets were constructed directly from the annotated genome and GFF files downloaded from NCBI RefSeq, ensuring comprehensive coverage of both canonical and non-canonical splice sites. Only forward-strand sequences were extracted, and redundant entries were removed.

The dataset is divided into two subsets: one exclusive for donor splice sites and the other for acceptor splice sites. Each subset contains a positive set of sequences with correct and biologically validated splice sites and a negative set of sequences that include the canonical dinucleotide motifs (`GT` for donor and `AG` for acceptor) but do not correspond to valid splice sites. For the negative sequences, the dinucleotides `GT`/`AG` appear at the same central positions (201–202) as in the positive sequences, preventing the model from relying solely on these motifs for discrimination. Table 1 describes the characteristics of the dataset, including the number of sequences in each subset.

To accurately identify U2-type introns, sequences were cross-referenced with the Intron Annotation and Orthology Database (IAOD) [36], which offers multi-species intron annotation. The combination of the DRANetSplicer dataset and the IAOD dataset provided a solid framework for a comprehensive analysis of U2-type introns, ensuring accurate identification of splice sites. Our analysis showed that the DRANetSplicer dataset contained 50,386 sequences with U2-type introns for the donor splice site dataset, while the acceptor dataset had 51,348 such sequences. After filtering the dataset for

Table 1 Characteristics of the DRANetSplicer dataset

	Donor splice sites	Acceptor splice sites
Sequence length	402	402
Number of positive samples	65,161	65,505
Number of negative samples	65,161	65,505

sequences containing U2-type introns, we explored two key hypotheses related to the effectiveness of splice site prediction models: (1) the impact of U2-type intron length and (2) the impact of the number of U2-type introns per sequence. Fig. 1 provides an overview of these hypotheses and the overall data description.

Hypotheses

Hypothesis 1: Short introns lead to higher effectiveness in splice site prediction

Our first hypothesis revolves around the potential improvement in the effectiveness of splice site prediction models when trained on a dataset containing short introns, as opposed to long ones. This hypothesis is grounded in the complex molecular dynamics and biochemical factors inherent to RNA splicing.

The length of introns varies widely among different organisms and different genes. Generally, introns in animal genes are longer than those in plant genes. For example, human genes tend to have short exons separated by long introns with a mean and median length of 3356 and 1023 bp, respectively, whereas introns in *Arabidopsis thaliana* have a mean and median length of 168 and 100 bp, respectively [9, 50]. Long introns tend to contain multiple splice sites, where splice sites compete for alternative splicing [19], and can introduce spatial complexities (for example, branch points located 50–60 nucleotides from the splice site; polypyrimidine tracts covering a broader sequence region) that obstruct the precise alignment of intron-exon junctions during the splicing process. These spatial complexities, in turn, increase the likelihood of activating cryptic splice sites [15], which are sequences that can trigger unexpected splicing events, including alternative splicing. Additionally, in cases with long flanking introns, the exon definition mechanism, which is the way the spliceosome identifies or recognizes exons to ensure that they are properly included in mature mRNA, becomes more dominant. This dominance presents a challenge to the spliceosome in accurately recognizing exon boundaries [44], as it relies on the precise identification of small exonic sequences, potentially resulting in exon-skipping events. Given this molecular background, we anticipate that splice site prediction models will demonstrate superior effectiveness when applied to a dataset containing short U2-type introns. Short U2-type introns inherently possess fewer complexities, making them a promising focus for improving the effectiveness of splice site prediction.

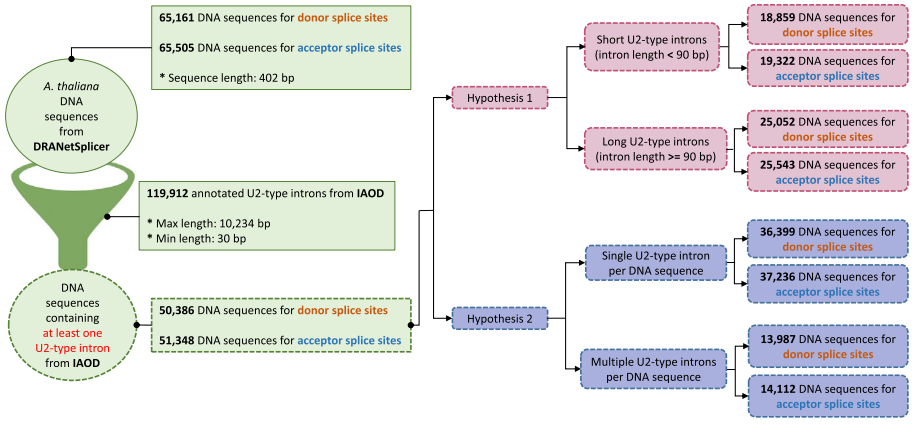

Fig. 1 Overview of the two hypotheses explored in this study, along with a description of the dataset

Within the DRANetSplicer dataset, we examined the U2-type intron length distribution. The donor set exhibited an average length of 99.6 bp, with a median length of 91 bp and ranged from 30 to 202 bp. Similarly, the acceptor set had an average U2-type intron length of 99.5 bp, with the same median length of 91 bp. The acceptor set had a similar range, with the length of U2-type introns varying from a minimum of 30 bp to a maximum of 202 bp. These observations demonstrate the significance of considering U2-type intron length when analyzing splice site prediction.

According to previous research, a mature intron, which consists of functional structure units that increase in number with intron length and are linked to form a cohesive structure that interacts with mRNA, is generally not shorter than 80 bp [55]. This same study classifies introns into short introns (ranging from 0 to 120 bp) and long introns (exceeding 120 bp), with an almost equal intron distribution over the two categories. Another study, specifically focusing on *Arabidopsis thaliana*, defines short introns as those shorter than 116 bp [27, 32] and notes that the proportion of short introns falls within the range of 45% to 65%. In our study, we define U2-type introns shorter than 90 bp as short U2-type introns. This choice is supported by the consistent median U2-type intron length of 91 bp in both our donor and acceptor datasets.

Based on the aforementioned classification criterion, the dataset included, on the one hand, 18,859 DNA sequences with short U2-type introns (37.4% of the positive samples) and 25,052 DNA sequences with long U2-type introns (49.7% of the positive samples) among the sequences containing donor splice sites. The remaining 6,475 DNA sequences (12.9%) contained a mix of short and long U2-type introns and were not used to test this hypothesis. On the other hand, there were 19,322 DNA sequences with short U2-type introns (37.6% of the positive samples) and 25,543 DNA sequences with long U2-type introns (49.7% of the positive samples) among the sequences containing acceptor splice sites. The remaining 6,483 DNA sequences (12.7%) contained a mix of short and long U2-type introns and were not used for testing this hypothesis.

To ensure a balanced and robust comparison, we organized the dataset into six distinct subsets: a donor set with short U2-type introns, a donor set with long U2-type introns, a donor set with an equal mixture of short and long U2-type introns, an acceptor set with short U2-type introns, an acceptor set with long U2-type introns, and an acceptor set with an equal mixture of short and long U2-type introns. We refer to the set containing an equal mixture of short and long U2-type introns as the length-mixed set. Given that the short intron category contained fewer samples compared to the long intron category, we randomly subsampled both categories to the same size for a fair comparison. This resulted in six working sets of 18,859 samples for the donor splice sites and 19,322 samples for the acceptor splice sites. By maintaining an identical number of sequences in each subset, potential biases were minimized, ensuring that any observed differences in prediction effectiveness could be attributed to the intrinsic properties of the sequences rather than the sample sizes.

We used the total number of negative samples (65,161 for donor and 65,505 for acceptor), resulting in a positive-to-negative ratio of approximately 1:3 for both donor and acceptor sites. This approach ensured a comprehensive representation of negative samples while maintaining sufficient positive samples for robust training. Finally, we randomly partitioned the data in each subset into 80% for training and 20% for testing.

Hypothesis 2: Multiple introns per sequence in a dataset lead to increased splice site prediction effectiveness

Our second hypothesis asserts that a dataset containing sequences with more than one U2-type intron will yield superior effectiveness in splice site prediction compared to a dataset containing sequences with only one U2-type intron. This hypothesis is built upon insights from splicing kinetics and the complex regulatory mechanisms controlling gene expression.

Specifically, kinetic analysis has indicated that the removal of the first intron in a sequence often improves the effectiveness with which subsequent splicing events occur. This improvement may arise from accelerated recruitment of the exon junction complex (EJC) to ligated exons or the establishment of a stable splicing framework that promotes the subsequent removal of introns [14].

Additionally, experimental results presented in [25] have shown that configurations of multiple introns within a transcript can effectively influence the splicing machinery. For example, full-length long-read sequencing techniques have revealed that in most cases, upstream introns tend to be spliced before downstream ones. This sequential splicing mechanism suggests that multi-intron genes may experience stronger evolutionary selection pressure for robust, consistently positioned splice site signals, since failure at any splicing step could result in non-functional transcripts [37]. If this selective pressure results in more canonical or stereotypical splice site sequences in multi-intron genes, these sequences would contain more consistent, learnable patterns for machine learning models. Consequently, this hypothesis anticipates that splice site prediction models will demonstrate improved effectiveness when trained on datasets containing multiple U2-type introns per sequence, as these sequences may harbor more robust and computationally recognizable splice site features [4].

We divided the sequences in the DRANetSplicer dataset into two categories: one with sequences containing only one U2-type intron and one with sequences containing more than one U2-type intron. Specifically, in the case of donor splice sites, the dataset consisted of 36,399 DNA sequences with a single U2-type intron per sequence and 13,987 DNA sequences with multiple U2-type introns per sequence. Similarly, for acceptor splice sites, the dataset included 37,236 DNA sequences with a single U2-type intron per sequence and 14,112 DNA sequences with multiple U2-type introns per sequence. We then organized the aforementioned two categories into six different subsets: a donor set with only one U2-type intron per sequence, a donor set with more than one U2-type intron per sequence, a donor set with an equal mixture of sequences with only one U2-type intron and sequences with more than one U2-type intron, an acceptor set with only one U2-type intron per sequence, an acceptor set with more than one U2-type intron per sequence, and an acceptor set with an equal mixture of sequences with only one U2-type intron and sequences with more than one U2-type intron. We refer to the set with an equal mixture of sequences with only one U2-type intron and sequences with more than one U2-type intron as count-mixed set.

As for the first hypothesis, since the multiple intron category contained fewer samples compared to the single intron category, we equalized the number of samples between the two for a fair comparison. We randomly subsampled each of these six subsets to obtain six corresponding working sets of 13,987 samples for the donor splice sites and 14,112 samples for the acceptor splice sites. By maintaining an identical number of

sequences in each subset, potential biases were minimized, ensuring that any observed differences in prediction effectiveness could be attributed to the intrinsic properties of the sequences rather than the size of the data set. We used the total number of negative samples (65,161 for donor and 65,505 for acceptor), resulting in a positive-to-negative ratio of approximately 1:4 for both donor and acceptor sites. Following this, we randomly partitioned each subset into 80% for training and 20% for testing. The partitioning was performed on disjoint sequence sets prior to cross-validation to ensure that no sequence appears in both training and test sets.

Model description

We built a novel CNN model named IntSplicer (Intron Splicer). To evaluate the effectiveness of our model, we benchmarked it against three existing models: SpliceRover [57], SpliceFinder [53], and DeepSplicer [1]. Each model has its unique architecture and approach to splice site prediction, thus making it possible to perform a comprehensive comparison. Table 2 describes the key characteristics of the different model architectures.

1. **IntSplicer**: Our model consists of multiple convolutional layers that play a crucial role in capturing essential sequence features, which are specific motifs that are crucial for identifying splice sites. The first layer consists of 64 filters with a kernel size of 10 and a stride of 4, followed by subsequent layers with increasing filter sizes (128, 256, and 512) and varying kernel sizes (3 and 2). Rectified Linear Unit (ReLU) activation functions are applied to introduce non-linearity. Between convolutional layers, max-pooling layers with a pool size of 2 are incorporated to down-sample the feature maps and reduce dimensionality. To prevent overfitting, dropout layers are placed after each max-pooling layer. Following the convolutional layers, a flatten layer transforms the feature maps into a one-dimensional vector. Next, a dense layer with 512 neurons and ReLU activation provides high-level abstractions of the extracted features. The final layer, consisting of two neurons with softmax activation, performs the

Table 2 Characteristics of the different model architectures

Layer	SpliceRover	SpliceFinder	DeepSplicer	IntSplicer
1	Conv 70 x (9, 4)	Conv 50 x (9, 4)	Conv 50 x (9, 4)	Conv 64 x (10, 4)
2	Dropout ($p = 0.2$)	Flatten	Conv 50 x (9, 1)	MaxPool (2, 1)
3	Conv 100 x (7, 1)	Dense 100 (ReLU)	Conv 50 x (9, 1)	Dropout ($p = 0.3$)
4	Dropout ($p = 0.2$)	Dropout ($p = 0.3$)	Flatten	Conv 128 x (3, 1)
5	Conv 100 x (7, 1)	Dense 2 (Softmax)	Dense 100 (ReLU)	MaxPool (2, 1)
6	MaxPool (4, 1)	–	Dropout ($p = 0.3$)	Dropout ($p = 0.2$)
7	Dropout ($p = 0.2$)	–	Dense 2 (Softmax)	Conv 256 x (3, 1)
8	Conv 200 x (7, 1)	–	–	MaxPool (2, 1)
9	MaxPool (4, 1)	–	–	Dropout ($p = 0.3$)
10	Dropout ($p = 0.2$)	–	–	Conv 512 x (2, 1)
11	Conv 250 x (7, 1)	–	–	MaxPool (2, 1)
12	MaxPool (4, 1)	–	–	Dropout ($p = 0.2$)
13	Dropout ($p = 0.2$)	–	–	Flatten
14	Flatten	–	–	Dense 512 (ReLU)
15	Dense 512 (ReLU)	–	–	Dropout ($p = 0.2$)
16	Dropout ($p = 0.2$)	–	–	Dense 2 (Softmax)
17	Dense 2 (Softmax)	–	–	-

binary classification for splice site prediction. IntSplicer, with its 1,962,946 trainable parameters, stands as a model of moderate complexity compared to its counterparts.

2. **SpliceRover**: SpliceRover focuses on human and *Arabidopsis thaliana* data. This CNN model uses convolutional layers with ReLU activations, dropout layers, and a final 512-unit dense layer leading to a softmax binary classification layer. The architecture and approach of this model represent an important baseline in terms of splice site classification effectiveness. SpliceRover has the highest number of trainable parameters (12,645,778) among the compared models. This indicates a considerably more complex model architecture, potentially allowing for more nuanced learning and pattern recognition capabilities. However, such complexity could also lead to challenges such as overfitting, especially if the training data are limited in terms of quantity and diversity. Compared to SpliceRover, IntSplicer is substantially less complex, with roughly 85% fewer trainable parameters. This could imply a more streamlined or focused learning approach, resulting in faster training times and reduced computational resource requirements [57].
3. **SpliceFinder**: SpliceFinder consists of a single convolutional layer, followed by one fully-connected layer and a softmax layer for classification. Its structure, emphasizing the capture of local sequence patterns and complex pattern processing, served as a key comparative model for evaluating the efficiency and effectiveness of our CNN model in splice site detection. The parameter count of SpliceFinder (2,012,152) is slightly higher than the parameter count of IntSplicer but still operates on a similar scale. The closeness in the number of parameters suggests that both models may have similar complexities and potentially similar capabilities in terms of learning and generalization [53].
4. **DeepSplicer**: DeepSplicer, another CNN-based architecture, includes three convolutional layers, alongside flatten, fully-connected, dropout, and softmax layers. DeepSplicer has a comparable number of trainable parameters (2,057,252) to SpliceFinder and slightly more than IntSplicer. This similarity in model size suggests that DeepSplicer and IntSplicer might have similar computational demands and generalization capabilities [1].

Experimental setup

We converted the DNA sequences into a numerical representation using one-hot encoding, where each nucleotide is represented as follows:

$$\mathrm{A} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \mathrm{C} = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \mathrm{G} = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \quad \mathrm{T} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}, \quad \mathrm{N} = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}$$

For training the models, we adopted a standardized approach to hyperparameter selection to ensure comparability and consistency across all models. This approach involved utilizing the Adam optimizer [29] with a learning rate of 0.001, the categorical cross-entropy loss function, and a batch size of 64, which empirically demonstrated optimal effectiveness in tests compared to batch sizes of 32, 128, and 256.

The Adam optimizer was selected for its effectiveness in handling sparse gradients and its adaptive learning rate properties, making it particularly well-suited in our study. The choice of a 0.001 learning rate was based on its widespread adoption as a value that

balances fast convergence with the stability of the training process across a variety of tasks and datasets.

The uniform application of the different hyperparameters across all models helped to facilitate a fair and direct comparison of their effectiveness, removing potential biases that could arise from different training conditions.

We applied a stratified 5-fold cross-validation approach to each of the 12 datasets (6 for each hypothesis) to ensure that models were trained and validated on data from the same intron category [48]. For each dataset, the 80% training portion was divided into 5 folds, with the model trained on 4 folds and validated on the remaining fold, repeating this process 5 times per dataset. Consequently, for each hypothesis (comprising 6 datasets), we conducted a total of $6 \times 5 = 30$ training sessions, ensuring that every model was both trained and validated within its specific intron category.

To further enhance the effectiveness of the training process and prevent overfitting, we used early stopping. This technique monitors the validation loss, stopping the training if no improvement was observed after a certain number of epochs, thereby ensuring that the models were not over-trained. We set the maximum number of epochs to 30 and used early stopping with patience 5, which means that training stops if no improvement is observed on the validation loss after 5 epochs.

Evaluation of effectiveness

In the context of splice site prediction, True Positives (TP) are splice sites correctly identified as either donor or acceptor sites. True Negatives (TN) are non-splice sites correctly identified as not being either donor or acceptor sites. False Positives (FP) are non-splice sites incorrectly identified as either donor or acceptor sites, and False Negatives (FN) are splice sites incorrectly identified as not being either donor or acceptor sites. With this understanding, we used four evaluation metrics to assess the prediction effectiveness of the models:

1. Sensitivity: Sensitivity, also known as recall, measures the proportion of actual positives correctly identified by the model. It is calculated as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{1}$$

2. Specificity: Specificity measures the proportion of actual negatives correctly identified by the model. It is calculated as:

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{2}$$

3. F1-score: The F1-score is a statistical measure used to evaluate the accuracy of a test. It considers both the precision (the proportion of true positive results in all positive predictions) and the recall (the proportion of true positive results over all actual positives). Specifically, the F1-score is the harmonic mean of precision and recall, calculated using the following formula:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{3}$$

4. Matthews Correlation Coefficient (MCC): The MCC is a correlation coefficient used to measure the quality of binary classifications. MCC is considered a balanced measure that can be used even if there is a large class imbalance. MCC is calculated using the following formula:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (4)$$

5. Area Under the Receiver Operating Characteristic Curve (AUROC): AUROC evaluates the model's ability to discriminate between positive and negative classes across all possible classification thresholds. A higher AUROC indicates that the model performs better in distinguishing true positives from false positives.
6. Area Under the Precision–Recall Curve (AUPRC): AUPRC summarizes the trade-off between precision and recall across thresholds, which is particularly informative in cases of class imbalance. A higher AUPRC value indicates that the model maintains high precision and recall even when the dataset is skewed toward one class.

The evaluation was performed on the test set of each of the 12 subsets, ensuring that our assessment reflected the ability of the models to generalize on unseen data. Since during training we applied 5-fold cross-validation for each of the subsets based on the two hypotheses, we calculated the sensitivity, specificity, F1-score, MCC, AUROC, and AUPRC on the test set. Then, we used their average and standard deviation to evaluate the effectiveness. All evaluations were performed using a fixed decision threshold of 0.5 across all models and cross-validation folds.

Sensitivity analysis of intron length effects

To validate the robustness of our length-based classification and examine the continuous relationship between intron length and prediction accuracy, we conducted a comprehensive sensitivity analysis across the full spectrum of intron lengths. We combined all positive sequences from short, long, and mixed length categories for both donor and acceptor splice sites, removed duplicates to create a unified positive dataset, and paired these with the complete negative dataset. This approach eliminated artificial categorical boundaries and enabled analysis across the natural distribution of intron lengths. For each positive sequence in the test set, we determined the corresponding intron length using *k*-mer-guided exact matching against our comprehensive U2-type intron database. We built *k*-mer indices to accelerate lookup and performed exact substring matching to assign lengths with high confidence. We used *k*-mer indices with $k = 12$, providing sufficient specificity for accurate intron sequence matching while maintaining sensitivity across the full range of intron lengths. We divided the detected intron lengths into deciles, creating ten bins each containing approximately 10% of the length-mapped positive samples. For each decile, we evaluated model performance using a consistent 1:3 positive:negative ratio, randomly sampling negative examples from the test set to maintain balanced evaluation conditions across length categories. We computed F1-scores and Matthews Correlation Coefficients for each decile with 95% bootstrap confidence intervals (200 iterations) to enable robust statistical comparison between length bins. We applied LOWESS smoothing ($fraction = 0.3$) to generate continuous performance

trends across the intron length spectrum, providing both discrete bin-based and smooth curve representations of the length-performance relationship.

Visualization

For visualization purposes, we used the saliency maps described in [47] to calculate the contribution scores of each nucleotide in a DNA sequence with a true positive splice site. This method involves calculating the gradients of the output of the model with respect to its input, known as saliency scores. After calculating the saliency scores, we normalized them for better visualization, as implemented in [57]. This normalization process can be mathematically represented as follows:

$$\text{Normalized Saliency Score} = \frac{x_{i,j}}{\frac{1}{100n}\sum_{i=1}^{n}\sum_{j=1}^{l}|x_{i,j}|}, \tag{5}$$

where $x_{i,j}$ represents the saliency score of the nucleotide at the j-th position of the i-th sequence, n is the total number of sequences, l is the sequence length, and the denominator computes the average of the absolute score sums across all sequences, scaled by 100.

For example, for the sequence `ACTG` with scores $(0.1, -0.2, 0.3, -0.05)$ and the sequence `CGTG` with scores $(0.5, 0.6, 0.15, -0.01)$, the absolute sums are 0.65 and 1.26, respectively. The average of these sums is then calculated, resulting in 0.955. This average is scaled down by a factor of 100, yielding a normalization constant $S_{\text{avg}} = 0.00955$. Using this constant, each score is normalized by dividing it by S_{avg}. The resulting normalized scores for the sequence `ACTG` are approximately $(10.47, -20.94, 31.41, -5.24)$ and for `CGTG` are $(52.36, 62.83, 15.71, -1.05)$. This approach ensures that the scores are adjusted relative to the average score magnitude across all sequences, enabling a consistent scale for comparison. The normalized scores are then associated with each nucleotide in the DNA sequence, creating a saliency map [30]. We used these normalized saliency scores to generate high-quality sequence logos using the Python package Logomaker [49].

All sequence positions in this study are reported using 0-based indexing, consistent with the internal representation used during model training and interpretation. Under this convention, the canonical splice site dinucleotides (`GT` for donors and `AG` for acceptors) occur at positions 200–201 in the sequence visualizations. To provide statistical validation of the biological interpretations derived from these saliency maps, we performed quantitative correlation analysis between model attention scores and predicted functional regions. Regional saliency scores were calculated by averaging saliency values within defined genomic regions based on established *Arabidopsis thaliana* splicing characteristics. For acceptor sites, we defined functional regions as follows: potential branch point region (positions 160–180, corresponding to 20–40 bp upstream), where mean saliency scores were calculated for adenine nucleotides (`A`) [51]; polypyrimidine tract region (positions 181–200), where mean saliency scores were calculated for pyrimidine nucleotides (`C` and `T`) [20]; and splice site region (positions 200–201). For donor sites, we analyzed upstream exonic consensus (`MAG`, positions 197–199), downstream intronic consensus (`RAGT`, positions 202–205), and extended donor consensus (`MAG|GTRAGT`, positions 196–205) based on the established dicot donor consensus sequence [21]. Pearson correlation coefficients were computed between regional scores and splice

site attention to test whether models systematically attend to biologically relevant features. GC content was calculated for each sequence to control for sequence composition effects. Statistical significance was evaluated using two-tailed tests ($p < 0.05$). This approach provides quantitative validation of visual saliency interpretations while controlling for potential sequence composition confounders.

Results

Effect of U2-type intron length

Table 3 summarizes the effectiveness of the different models on the test data containing short, long, and length-mixed U2-type introns.

For donor splice sites, the models demonstrated higher effectiveness for short U2-type introns, with IntSplicer achieving the highest F1-score (0.940 ± 0.002) and MCC value (0.923 ± 0.002) in the short intron category. In contrast, the long and length-mixed intron categories showed slightly reduced effectiveness across all models, with F1-scores typically decreasing by 0.004 to 0.008. However, SpliceRover maintained comparable performance in the long intron category, achieving an F1-score of 0.939 ± 0.002 and an MCC value of 0.923 ± 0.003, which were the highest overall in that category.

For acceptor splice sites, the impact of intron length was more pronounced. The models showed notably higher effectiveness for the short intron category, with IntSplicer achieving an optimal F1-score of 0.933 ± 0.002 and an MCC value of 0.913 ± 0.002. The effectiveness decreased considerably for the long and length-mixed intron categories, with MCC values dropping by approximately 0.025 to 0.030 for all models.

Overall, IntSplicer demonstrated the highest effectiveness, particularly for short U2-type intron prediction, followed closely by SpliceRover. Both models consistently outperformed SpliceFinder and DeepSplicer across all categories for both donor and acceptor splice site prediction.

Detailed sensitivity analysis by intron length

Our sensitivity analysis revealed that the relationship between intron length and prediction accuracy follows a bell curve rather than the simple binary classification we initially hypothesized (Fig. 2, Fig. 3). Peak performance occurred at 87–90 bp for acceptor sites, closely validating our original 90 bp threshold choice.

Analysis of prediction accuracy across ten U2-type intron length deciles reveals fundamentally different sensitivities for donor and acceptor splice sites, as detailed in Table 4. All metrics are quoted from the best-performing model for that specific bin (SpliceRover generally, with IntSplicer providing the higher performance for the shortest acceptor bin).

Donor sites consistently demonstrated highly stable performance across all intron length deciles, supporting the finding that donor site prediction is less sensitive to intron length than acceptor site prediction. The F1-score, for instance, ranges narrowly from a minimum of 0.918 ($30 - 78$ bp) to a maximum of 0.955 ($86 - 89$ bp) using the SpliceRover model. The MCC metrics show a similarly consistent pattern of high accuracy (ranging from 0.893 to 0.939).

In contrast, acceptor sites exhibited a pronounced length-dependent performance profile, characterized by distinct tiers of prediction accuracy. Prediction accuracy is lowest for the shortest and longest introns and peaks sharply in the $82 - 93$ bp range.

Table 3 Comparison of short, long, and length-mixed intron categories for each model. Bold values indicate the highest performance scores achieved for each metric across the compared categories

Model	Metric	Short	Long	Length-mixed
Donor sites				
IntSplicer	Sensitivity	0.920 ± 0.008	**0.923 ± 0.007**	0.919 ± 0.010
	Specificity	**0.989 ± 0.002**	0.986 ± 0.003	0.985 ± 0.002
	F1-score	**0.940 ± 0.002**	0.936 ± 0.002	0.932 ± 0.003
	MCC	**0.923 ± 0.002**	0.918 ± 0.002	0.913 ± 0.003
	AUROC	**0.994 ± 0.001**	0.992 ± 0.000	0.992 ± 0.001
	AUPRC	**0.983 ± 0.002**	0.981 ± 0.000	0.980 ± 0.001
SpliceRover	Sensitivity	0.912 ± 0.007	**0.922 ± 0.008**	0.904 ± 0.004
	Specificity	**0.990 ± 0.001**	0.988 ± 0.003	0.988 ± 0.001
	F1-score	0.938 ± 0.003	**0.939 ± 0.002**	0.930 ± 0.002
	MCC	0.921 ± 0.003	**0.923 ± 0.003**	0.911 ± 0.002
	AUROC	**0.993 ± 0.001**	0.991 ± 0.001	0.990 ± 0.001
	AUPRC	**0.983 ± 0.001**	0.980 ± 0.001	0.970 ± 0.002
SpliceFinder	Sensitivity	0.908 ± 0.009	**0.914 ± 0.006**	0.904 ± 0.017
	Specificity	**0.984 ± 0.002**	0.981 ± 0.002	0.980 ± 0.005
	F1-score	**0.924 ± 0.002**	**0.924 ± 0.002**	0.917 ± 0.002
	MCC	**0.903 ± 0.002**	0.902 ± 0.002	0.894 ± 0.003
	AUROC	**0.994 ± 0.002**	0.990 ± 0.001	0.993 ± 0.002
	AUPRC	**0.982 ± 0.003**	0.972 ± 0.002	0.978 ± 0.003
DeepSplicer	Sensitivity	**0.927 ± 0.008**	0.924 ± 0.006	0.921 ± 0.017
	Specificity	**0.982 ± 0.005**	0.982 ± 0.003	0.982 ± 0.006
	F1-score	**0.933 ± 0.003**	0.930 ± 0.004	0.928 ± 0.003
	MCC	**0.914 ± 0.005**	0.910 ± 0.005	0.908 ± 0.003
	AUROC	**0.996 ± 0.001**	0.991 ± 0.001	0.994 ± 0.002
	AUPRC	**0.987 ± 0.002**	0.978 ± 0.002	0.982 ± 0.005
Acceptor sites				
IntSplicer	Sensitivity	**0.923 ± 0.006**	0.909 ± 0.016	0.905 ± 0.010
	Specificity	**0.984 ± 0.003**	0.976 ± 0.005	0.975 ± 0.003
	F1-score	**0.933 ± 0.002**	0.913 ± 0.003	0.909 ± 0.001
	MCC	**0.913 ± 0.002**	0.888 ± 0.003	0.883 ± 0.002
	AUROC	**0.993 ± 0.001**	0.987 ± 0.000	0.991 ± 0.001
	AUPRC	**0.981 ± 0.001**	0.964 ± 0.001	0.972 ± 0.002
SpliceRover	Sensitivity	**0.916 ± 0.009**	0.911 ± 0.010	0.893 ± 0.010
	Specificity	**0.986 ± 0.002**	0.977 ± 0.004	0.978 ± 0.003
	F1-score	**0.932 ± 0.002**	0.917 ± 0.002	0.907 ± 0.002
	MCC	**0.913 ± 0.002**	0.892 ± 0.003	0.881 ± 0.002
	AUROC	**0.994 ± 0.001**	0.986 ± 0.001	0.990 ± 0.001
	AUPRC	**0.985 ± 0.001**	0.964 ± 0.002	0.970 ± 0.002
SpliceFinder	Sensitivity	**0.902 ± 0.006**	0.892 ± 0.022	0.892 ± 0.006
	Specificity	**0.974 ± 0.002**	0.968 ± 0.009	0.969 ± 0.004
	F1-score	**0.906 ± 0.002**	0.892 ± 0.002	0.893 ± 0.004
	MCC	**0.878 ± 0.002**	0.860 ± 0.003	0.862 ± 0.005
	AUROC	**0.992 ± 0.002**	0.984 ± 0.001	0.987 ± 0.002
	AUPRC	**0.976 ± 0.005**	0.955 ± 0.001	0.963 ± 0.005
DeepSplicer	Sensitivity	**0.909 ± 0.009**	0.902 ± 0.023	0.902 ± 0.013
	Specificity	**0.982 ± 0.006**	0.976 ± 0.008	0.975 ± 0.006
	F1-score	**0.922 ± 0.006**	0.910 ± 0.005	0.908 ± 0.004
	MCC	**0.900 ± 0.008**	0.884 ± 0.006	0.882 ± 0.005
	AUROC	**0.993 ± 0.001**	0.988 ± 0.001	0.993 ± 0.002
	AUPRC	**0.981 ± 0.002**	0.969 ± 0.002	0.978 ± 0.003

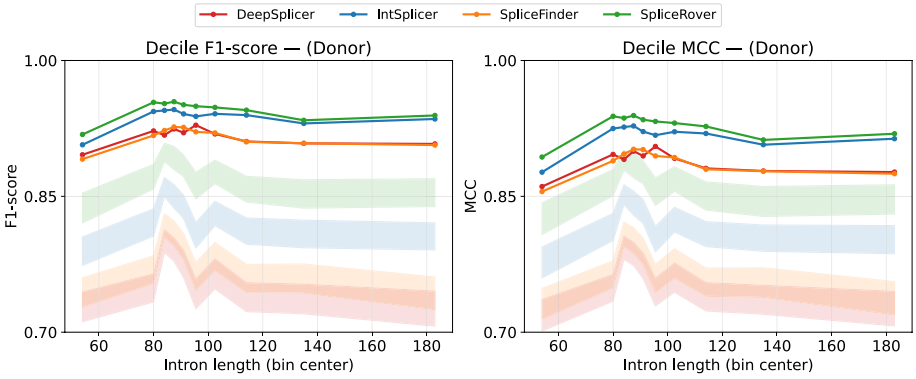

Fig. 2 Decile performance versus intron length for donor sites. Left: F1-score; right: MCC. Curves show IntSplicer, SpliceRover, SpliceFinder, and DeepSplicer; markers denote bin centers. Shaded bands indicate 95% bootstrap confidence intervals

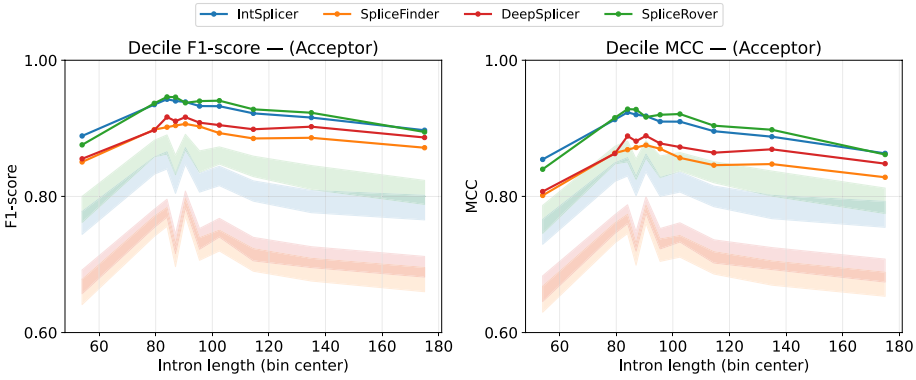

Fig. 3 Decile performance versus intron length for acceptor sites. Left: F1-score; right: MCC. Curves show IntSplicer, SpliceRover, SpliceFinder, and DeepSplicer; markers denote bin centers. Shaded bands indicate 95% bootstrap confidence intervals

Specifically, performance is significantly reduced for very short introns (< 77 bp) at F1 -score$= 0.889$ and MCC $= 0.854$ (IntSplicer metric). Performance then peaks in the optimal range ($82 - 86$ bp) reaching F1-score$= 0.946$ and MCC $= 0.928$ (SpliceRover max decile), followed by a gradual decline for longer introns, eventually reaching F1 -score$= 0.895$ and MCC $= 0.862$ for the longest introns (> 148 bp). Bootstrap confidence intervals confirm statistically significant differences between these performance tiers.

The continuous analysis of the smoothed trends demonstrates that while the choice of 90 bp as a binary classification cutoff generally captured the transition from optimal to declining performance, the underlying relationship is a smooth curve rather than a sharp boundary. This finding strengthens the hypothesis that splice site prediction accuracy is fundamentally linked to intron length while also revealing the detailed biological complexity involved.

Effect of U2-type intron count

Following the analysis of intron length impact, we examined the effect of U2-type intron count per sequence on model effectiveness. Table 5 summarizes our test results.

Table 4 Performance metrics (F1-score and MCC) across all intron length decile bins for best models

Intron length range (bp)	F1-score	MCC Score	Model
Donor sites (SpliceRover)			
$30-78$	0.918	0.893	SpliceRover
$78-82$	0.954	0.938	SpliceRover
$82-86$	0.952	0.936	SpliceRover
$86-89$	0.955	0.939	SpliceRover (Max performance)
$89-93$	0.951	0.935	SpliceRover
$93-98$	0.950	0.933	SpliceRover
$98-107$	0.948	0.931	SpliceRover
$107-122$	0.945	0.927	SpliceRover
$122-148$	0.934	0.912	SpliceRover
$148-202$	0.939	0.919	SpliceRover
Acceptor sites (SpliceRover, IntSplicer* for first bin)			
$31-77$	0.889	0.854	IntSplicer* (Reduced tier)
$77-82$	0.937	0.916	SpliceRover
$82-86$	0.946	0.928	SpliceRover (Optimal tier)
$86-88$	0.946	0.928	SpliceRover
$88-93$	0.938	0.917	SpliceRover
$93-98$	0.940	0.920	SpliceRover
$98-107$	0.941	0.921	SpliceRover
$107-122$	0.928	0.904	SpliceRover
$122-148$	0.923	0.898	SpliceRover
$148-202$	0.895	0.862	SpliceRover (Declining tier)

For donor splice sites, IntSplicer and SpliceRover showed particularly strong performance when tested on sequences containing multiple U2-type introns, achieving the highest F1-scores (0.950 ± 0.001 and 0.948 ± 0.002, respectively) and MCC values (0.940 ± 0.002 and 0.937 ± 0.003, respectively). The models consistently showed lower effectiveness on sequences containing single or count-mixed U2-type introns; for instance, F1-scores typically decreased by 0.020 to 0.025.

For acceptor splice sites, the impact of the U2-type intron count was more pronounced. The highest predictive performance was observed in sequences containing multiple U2-type introns; for example, IntSplicer and SpliceRover achieved F1-scores of 0.933 ± 0.004 and 0.932 ± 0.002, respectively, along with MCC values of 0.919 ± 0.005 and 0.917 ± 0.002. The performance gap was larger for acceptor splice sites, with F1-scores dropping by approximately 0.040 to 0.050 when testing on sequences containing a single U2-type intron each.

Overall, IntSplicer demonstrated the highest effectiveness across all sequence types, followed closely by SpliceRover, with both models consistently outperforming SpliceFinder and DeepSplicer. The results clearly indicate that the models achieved better effectiveness for both donor and acceptor splice site prediction for sequences containing multiple U2-type introns.

Cross-evaluation between single and multiple U2-type intron contexts

To further assess model robustness and context generalization, we performed cross-evaluation experiments where models trained on single U2-type intron sequences were tested on multiple U2-type intron sequences and vice versa (Table 6). Overall, within-category evaluations (*Single?Single* and *Multiple?Multiple*) yielded consistently strong performance across all architectures, with AUROC and AUPRC values

Table 5 Comparison of single, multiple, and count-mixed intron categories for each model. Bold values indicate the highest performance scores achieved for each metric across the compared categories

Model	Metric	Single	Multiple	Count-mixed
Donor sites				
IntSplicer	Sensitivity	0.931 ± 0.009	**0.941 ± 0.009**	0.921 ± 0.010
	Specificity	0.985 ± 0.003	**0.992 ± 0.002**	0.987 ± 0.002
	F1-score	0.930 ± 0.003	**0.950 ± 0.001**	0.923 ± 0.002
	MCC	0.915 ± 0.003	**0.940 ± 0.002**	0.914 ± 0.003
	AUROC	0.987 ± 0.000	**0.997 ± 0.001**	0.995 ± 0.000
	AUPRC	0.967 ± 0.001	**0.989 ± 0.001**	0.982 ± 0.001
SpliceRover	Sensitivity	0.912 ± 0.012	**0.933 ± 0.009**	0.915 ± 0.009
	Specificity	0.988 ± 0.002	**0.992 ± 0.002**	0.989 ± 0.002
	F1-score	0.926 ± 0.004	**0.948 ± 0.002**	0.931 ± 0.002
	MCC	0.911 ± 0.004	**0.937 ± 0.003**	0.917 ± 0.002
	AUROC	0.985 ± 0.001	**0.997 ± 0.001**	0.995 ± 0.000
	AUPRC	0.965 ± 0.001	**0.989 ± 0.002**	0.983 ± 0.001
SpliceFinder	Sensitivity	0.897 ± 0.026	**0.918 ± 0.009**	0.897 ± 0.012
	Specificity	0.983 ± 0.007	**0.986 ± 0.002**	0.982 ± 0.003
	F1-score	0.909 ± 0.005	**0.926 ± 0.003**	0.906 ± 0.004
	MCC	0.890 ± 0.006	**0.910 ± 0.004**	0.886 ± 0.004
	AUROC	0.987 ± 0.001	**0.997 ± 0.001**	0.995 ± 0.001
	AUPRC	0.961 ± 0.001	**0.986 ± 0.003**	0.978 ± 0.003
DeepSplicer	Sensitivity	0.925 ± 0.019	**0.939 ± 0.013**	0.911 ± 0.011
	Specificity	0.980 ± 0.004	**0.987 ± 0.005**	0.984 ± 0.003
	F1-score	0.917 ± 0.003	**0.940 ± 0.007**	0.917 ± 0.005
	MCC	0.900 ± 0.003	**0.927 ± 0.008**	0.899 ± 0.006
	AUROC	0.987 ± 0.001	**0.998 ± 0.001**	0.996 ± 0.001
	AUPRC	0.963 ± 0.002	**0.991 ± 0.002**	0.981 ± 0.002
Acceptor sites				
IntSplicer	Sensitivity	0.879 ± 0.018	**0.934 ± 0.004**	0.885 ± 0.014
	Specificity	0.980 ± 0.004	**0.985 ± 0.003**	0.982 ± 0.003
	F1-score	0.890 ± 0.003	**0.933 ± 0.004**	0.900 ± 0.004
	MCC	0.868 ± 0.003	**0.919 ± 0.005**	0.879 ± 0.004
	AUROC	0.984 ± 0.001	**0.996 ± 0.001**	0.992 ± 0.001
	AUPRC	0.950 ± 0.001	**0.985 ± 0.002**	0.970 ± 0.002
SpliceRover	Sensitivity	0.859 ± 0.015	**0.924 ± 0.007**	0.878 ± 0.012
	Specificity	0.981 ± 0.004	**0.987 ± 0.002**	0.980 ± 0.003
	F1-score	0.883 ± 0.002	**0.932 ± 0.002**	0.891 ± 0.004
	MCC	0.860 ± 0.002	**0.917 ± 0.002**	0.868 ± 0.005
	AUROC	0.984 ± 0.001	**0.996 ± 0.001**	0.992 ± 0.001
	AUPRC	0.950 ± 0.002	**0.986 ± 0.002**	0.970 ± 0.003
SpliceFinder	Sensitivity	0.868 ± 0.019	**0.921 ± 0.011**	0.861 ± 0.029
	Specificity	0.973 ± 0.005	0.968 ± 0.004	**0.977 ± 0.005**
	F1-score	0.870 ± 0.003	**0.891 ± 0.004**	0.876 ± 0.007
	MCC	0.843 ± 0.003	**0.867 ± 0.005**	0.850 ± 0.006
	AUROC	0.983 ± 0.000	**0.994 ± 0.001**	0.991 ± 0.002
	AUPRC	0.940 ± 0.001	**0.970 ± 0.003**	0.957 ± 0.005
DeepSplicer	Sensitivity	0.885 ± 0.015	**0.918 ± 0.019**	0.903 ± 0.018
	Specificity	0.978 ± 0.004	**0.983 ± 0.003**	0.976 ± 0.005
	F1-score	0.891 ± 0.004	**0.918 ± 0.005**	0.897 ± 0.006
	MCC	0.868 ± 0.004	**0.901 ± 0.006**	0.875 ± 0.008
	AUROC	0.985 ± 0.001	**0.997 ± 0.001**	0.993 ± 0.003
	AUPRC	0.951 ± 0.003	**0.986 ± 0.002**	0.969 ± 0.008

Table 6 Comparison of within-category and cross-evaluation results for each model. Columns show training→testing settings. Bold values indicate the highest score per metric across the four evaluation settings

Model	Metric	Single→Single	Multiple→Multiple	Single→Multiple	Multiple→Single
Donor sites					
IntSplicer	Sensitivity	0.931 ± 0.009	**0.941 ± 0.009**	0.914 ± 0.014	0.781 ± 0.016
	Specificity	0.985 ± 0.003	0.992 ± 0.002	0.992 ± 0.002	**0.993 ± 0.001**
	F1-score	0.930 ± 0.003	0.950 ± 0.001	**0.951 ± 0.007**	0.873 ± 0.011
	MCC	0.915 ± 0.003	**0.940 ± 0.002**	0.908 ± 0.012	0.792 ± 0.015
	AUROC	0.987 ± 0.000	**0.997 ± 0.001**	0.992 ± 0.000	0.982 ± 0.002
	AUPRC	0.967 ± 0.001	0.989 ± 0.001	**0.993 ± 0.000**	0.985 ± 0.002
SpliceRover	Sensitivity	0.912 ± 0.012	**0.933 ± 0.009**	0.905 ± 0.019	0.768 ± 0.029
	Specificity	0.988 ± 0.002	0.992 ± 0.002	0.992 ± 0.002	**0.995 ± 0.002**
	F1-score	0.926 ± 0.004	**0.948 ± 0.002**	0.946 ± 0.010	0.866 ± 0.018
	MCC	0.911 ± 0.004	**0.937 ± 0.003**	0.901 ± 0.017	0.783 ± 0.023
	AUROC	0.985 ± 0.001	**0.997 ± 0.001**	0.992 ± 0.001	0.982 ± 0.002
	AUPRC	0.965 ± 0.001	0.989 ± 0.002	**0.993 ± 0.001**	0.985 ± 0.002
SpliceFinder	Sensitivity	0.897 ± 0.026	**0.918 ± 0.009**	0.877 ± 0.036	0.774 ± 0.025
	Specificity	0.983 ± 0.007	0.986 ± 0.002	0.992 ± 0.005	**0.995 ± 0.002**
	F1-score	0.909 ± 0.005	0.926 ± 0.003	**0.930 ± 0.019**	0.870 ± 0.016
	MCC	0.890 ± 0.006	**0.910 ± 0.004**	0.875 ± 0.029	0.789 ± 0.022
	AUROC	0.987 ± 0.001	**0.997 ± 0.001**	0.990 ± 0.001	0.983 ± 0.001
	AUPRC	0.961 ± 0.001	0.986 ± 0.003	**0.992± 0.001**	0.975 ± 0.001
DeepSplicer	Sensitivity	0.925 ± 0.019	**0.939 ± 0.013**	0.905 ± 0.025	0.789 ± 0.048
	Specificity	0.980 ± 0.004	0.987 ± 0.005	0.988 ± 0.004	**0.994 ± 0.003**
	F1-score	0.917 ± 0.003	0.940 ± 0.007	**0.944 ± 0.013**	0.878 ± 0.029
	MCC	0.900 ± 0.003	**0.927 ± 0.008**	0.897 ± 0.020	0.800 ± 0.038
	AUROC	0.987 ± 0.001	**0.998 ± 0.001**	0.991 ± 0.001	0.983 ± 0.001
	AUPRC	0.963 ± 0.002	0.991 ± 0.002	**0.992 ± 0.001**	0.978 ± 0.002
Acceptor sites					
IntSplicer	Sensitivity	0.879 ± 0.018	**0.934 ± 0.004**	0.863 ± 0.016	0.676 ± 0.027
	Specificity	0.980 ± 0.004	0.985 ± 0.003	0.987 ± 0.002	**0.990 ± 0.002**
	F1-score	0.890 ± 0.003	**0.933 ± 0.004**	0.920 ± 0.008	0.802 ± 0.019
	MCC	0.868 ± 0.003	**0.919 ± 0.005**	0.857 ± 0.013	0.702 ± 0.021
	AUROC	0.984 ± 0.001	**0.996±0.001**	0.986 ± 0.001	0.973 ± 0.002
	AUPRC	0.950 ± 0.001	0.985 ± 0.002	**0.987 ± 0.001**	0.975 ± 0.002
SpliceRover	Sensitivity	0.859 ± 0.015	**0.924 ± 0.007**	0.833 ± 0.015	0.611 ± 0.086
	Specificity	0.981 ± 0.004	0.987 ± 0.002	0.988 ± 0.001	**0.994 ± 0.003**
	F1-score	0.883 ± 0.002	**0.932 ± 0.002**	0.903 ± 0.009	0.753 ± 0.066
	MCC	0.860 ± 0.002	**0.917 ± 0.002**	0.831 ± 0.012	0.655 ± 0.064
	AUROC	0.984 ± 0.001	**0.996 ± 0.001**	0.984 ± 0.001	0.971 ± 0.001
	AUPRC	0.950 ± 0.002	**0.986 ± 0.002**	0.985 ± 0.001	0.973 ± 0.001
SpliceFinder	Sensitivity	0.868 ± 0.019	**0.921 ± 0.011**	0.804 ± 0.032	0.775 ± 0.029
	Specificity	0.973 ± 0.005	0.968 ± 0.004	**0.989 ± 0.003**	0.984 ± 0.003
	F1-score	0.870 ± 0.003	**0.891 ± 0.004**	0.885 ± 0.019	0.866 ± 0.017
	MCC	0.843 ± 0.003	**0.867 ± 0.005**	0.806 ± 0.025	0.777 ± 0.022
	AUROC	0.983 ± 0.000	**0.994 ± 0.001**	0.982 ± 0.001	0.972 ± 0.002
	AUPRC	0.940 ± 0.001	0.970 ± 0.003	**0.983 ± 0.001**	0.975 ± 0.002
DeepSplicer	Sensitivity	0.885 ± 0.015	**0.918 ± 0.019**	0.852 ± 0.034	0.699 ± 0.060
	Specificity	0.978 ± 0.004	0.983 ± 0.003	0.987 ± 0.006	**0.992 ± 0.003**
	F1-score	0.891 ± 0.004	**0.918 ± 0.005**	0.913 ± 0.018	0.818 ± 0.041
	MCC	0.868 ± 0.004	**0.901 ± 0.006**	0.847 ± 0.025	0.723 ± 0.046
	AUROC	0.985 ± 0.001	**0.997 ± 0.001**	0.986 ± 0.001	0.975 ± 0.002
	AUPRC	0.951 ± 0.003	0.986 ± 0.002	**0.987 ± 0.001**	0.978 ± 0.002

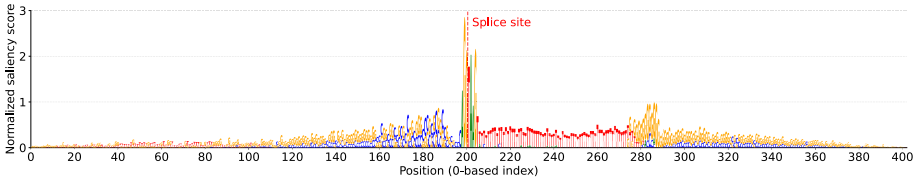

Fig. 4 Sequence logo for donor sequences containing short U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide GT shown at positions 200–201 using 0-based indexing.

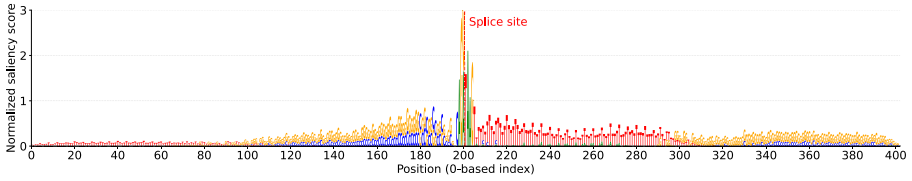

Fig. 5 Sequence logo for donor sequences containing long U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide GT shown at positions 200–201 using 0-based indexing.

typically above 0.98. In cross-evaluation, performance generally decreased, particularly in the *Multiple?Single* setting. For example, IntSplicer retained a high donor F1-score of 0.951 ± 0.007 in the *Single?Multiple* evaluation, but dropped to 0.873 ± 0.011 in the *Multiple?Single* case, while SpliceRover showed a similar decline from 0.946 ± 0.010 to 0.866 ± 0.018. For acceptor sites, the same pattern was observed, with IntSplicer F1-scores decreasing from 0.920 ± 0.008 in the *Single?Multiple* setting to 0.802 ± 0.019 in the *Multiple?Single* evaluation. These results indicate that models trained on single U2-type intron sequences generalize reasonably well compared to multiple U2-type intron sequences, whereas models trained on U2-type intron sequences show weaker generalization to single U2-type intron sequences, likely due to the greater contextual variability and auxiliary motif density in multi-intron genes.

Discussion

Models trained on datasets consisting of short U2-type introns demonstrate a higher predictive effectiveness for acceptor splice sites compared to those trained on datasets with long U2-type introns, whereas no substantial difference is observed for donor splice sites

The first hypothesis suggested that models trained on datasets with short U2-type introns would outperform those trained on datasets with long U2-type introns in predicting both donor and acceptor splice sites. Our experimental results confirmed this hypothesis for the prediction of acceptor splice sites for all models. However, we did not observe a substantial difference for donor splice sites.

The sequence logos (Figs. 4 to 7) provide visual evidence to support these observations. For donor splice sites, the sequence logos for short U2-type introns (Fig. 4) and long U2-type introns (Fig. 5) show similar core patterns, characterized by consistent G and T signals at positions 200 and 201, which correspond to the donor splice site in the dataset.

The downstream polypyrimidine tract is visibly shorter in the sequence logo for short introns, whereas it extends further for long introns, making the difference in intron length visually apparent. However, both logos show comparable peak saliency scores

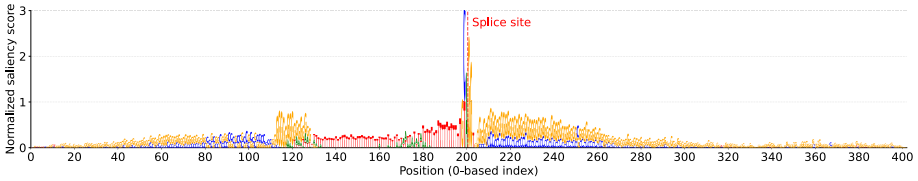

Fig. 6 Sequence logo for acceptor sequences containing short U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide AG shown at positions 200–201 using 0-based indexing

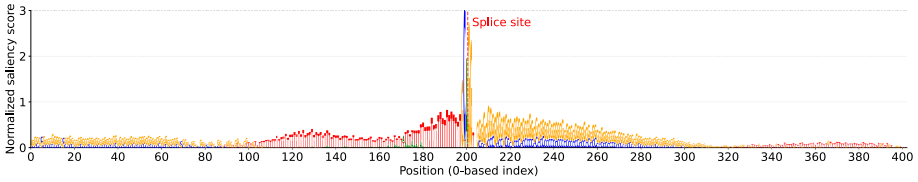

Fig. 7 Sequence logo for acceptor sequences containing long U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide AG shown at positions 200–201 using 0-based indexing.

near the donor site, indicating that the model relies on essentially the same key signals for donor site recognition in both short and long introns. As a result, the difference in polypyrimidine tract length does not translate into a clear predictive advantage for donor sites in short introns.

For acceptor splice sites, the sequence logos for short U2-type introns (Fig. 6) and long U2-type introns (Fig. 7) display the AG dinucleotide at positions 200 and 201, which correspond to the acceptor splice site in the dataset. The saliency scores of the branch point signal (A around positions 160–180) and the polypyrimidine tract are higher in Fig. 6 compared to Fig. 7. Similarly to donor splice sites, the span of the polypyrimidine tract extends less than 90 bp in Fig. 6 but stretches beyond 90 bp in Fig. 7, providing a visual indication of the importance of intron length. The position of the branch point, reflecting the typical 18–40 bp distance in *Arabidopsis thaliana* [6], and its higher saliency score in Fig. 6 aligns with the observations in [24], supporting a more uniform splicing mechanism. These higher saliency scores observed in key splicing elements help explain the improved prediction effectiveness for acceptor splice sites in short U2-type introns.

Furthermore, long introns could also contain more splicing regulatory elements such as enhancers or silencers, which modulate the kinetics and efficiency of spliceosome assembly at nearby splice sites, thereby influencing splicing outcomes. These intron elements are important for adjusting where splicing takes place. Furthermore, according to [12, 23], shorter introns can lead to more effective processing and recognition of the splicing machinery, which in turn could facilitate more consistent regulation of gene expression.

Quantitative correlation analysis provided statistical validation of the visual patterns observed in the sequence logos. For donor splice sites, the correlations between short U2-type introns (Fig. 8) and long U2-type introns confirmed the similar core recognition patterns observed in the sequence logos. The extended donor consensus (MAG | GTRAGT) showed exceptionally strong correlations with splice site scores for both short (r=0.90) and long (r=0.88) introns, while individual upstream MAG and downstream RAGT components exhibited much weaker correlations (r=0.24−0.16 and r=0.14−0.08, respectively). This demonstrates that models recognize donor sites through integrated

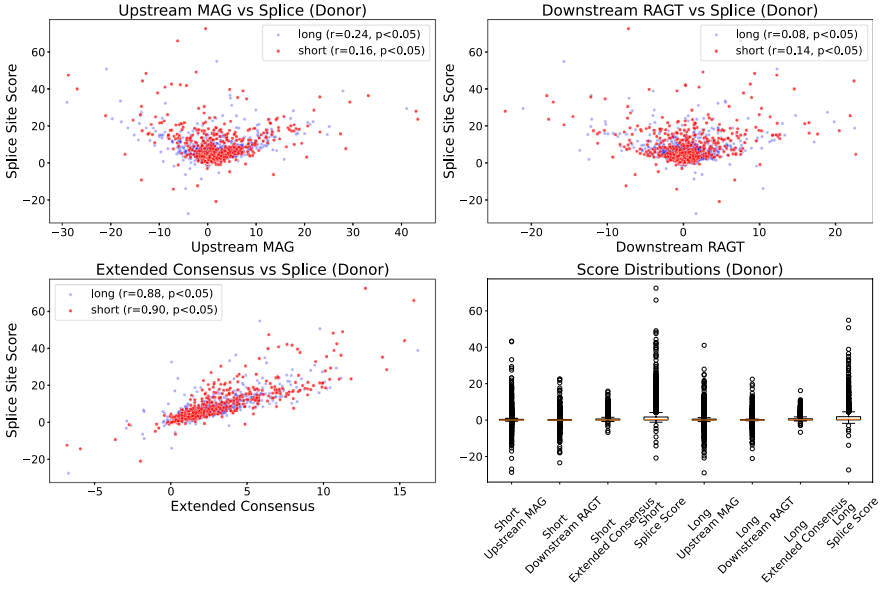

Fig. 8 Quantitative validation of model attention patterns for donor sites (short vs long). Colors indicate intron category: short (red) and long (blue). Each panel shows the relationship between regional normalized saliency scores and the splice-site score. The legend reports the Pearson correlation coefficient (r) and the two-sided p-value; $p<0.05$ denotes a statistically significant correlation at the 5% level, and $p\geq 0.05$ denotes a non-significant result. Top left: upstream MAG consensus vs splice score. Top right: downstream RAGT consensus vs splice score. Bottom left: extended donor consensus (MAG GTRAGT) vs splice score. Bottom right: distributions of regional scores

consensus sequence analysis rather than individual motif components, supporting the visual observation of comparable peak saliency scores near donor sites regardless of intron length. For acceptor splice sites, the quantitative analysis (Fig. 9) validated the visual differences observed between short and long U2-type introns. Polypyrimidine tract correlations remained consistently strong for both short (r=0.60) and long (r=0.59) introns, while branch point correlations showed moderate values (r=0.21 and r=0.27, respectively). The minimal GC content correlations (r=−0.10 to −0.08) confirmed that observed patterns reflect genuine biological signal recognition rather than sequence composition artifacts. These results statistically support the visual observation of higher saliency scores in key splicing elements for short introns.

The major observations and biological considerations are: (1) Shorter introns may concentrate essential splicing signals (e.g., branch points and polypyrimidine tracts) within a smaller region, which appears to correlate with more consistent acceptor site recognition in the models. (2) The conserved 18–40 bp distance between the branch point and the acceptor site in *Arabidopsis thaliana* provides a stable and reliable feature for model prediction, particularly in shorter introns where this distance is consistently maintained. (3) The variable distance between the branch point and donor sites, which tends to expand with increasing intron length, could explain why short introns do not lead to higher predictive performance for donor site recognition compared with acceptor site prediction. (4) The proximity of the branch point and the polypyrimidine tract to the acceptor site provides an additional predictive cue; in shorter introns, these signals are more localized, which likely contributes to the higher predictive accuracy observed for acceptor site prediction.

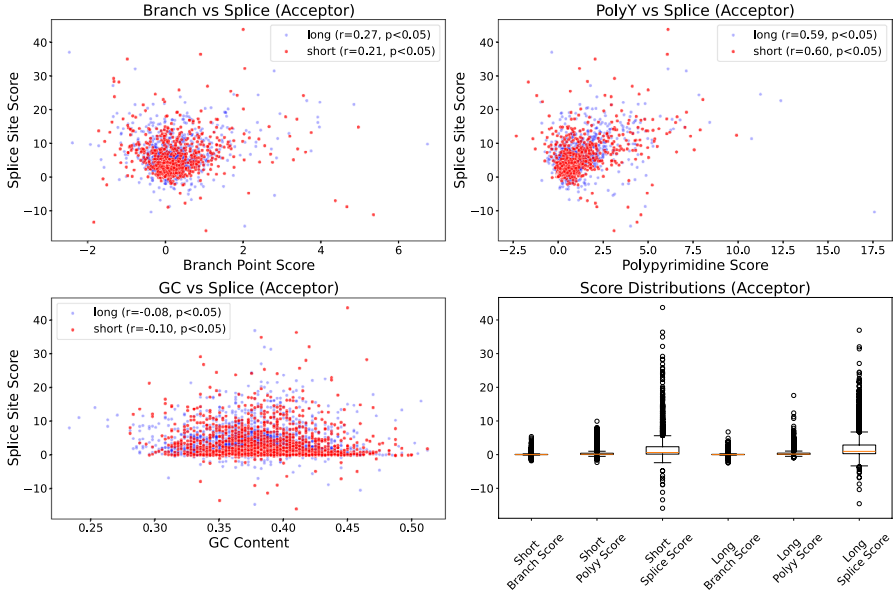

Fig. 9 Quantitative validation of model attention patterns for acceptor sites (short vs long). Colors indicate intron category: short (red) and long (blue). Each panel shows the relationship between regional normalized saliency scores and the splice-site score. The legend reports the Pearson correlation coefficient (*r*) and the two-sided *p*-value; $p<0.05$ denotes a statistically significant correlation at the 5% level, and $p\geq 0.05$ denotes a non-significant result. Top left: branch point adenine region vs splice score. Top right: polypyrimidine tract vs splice score. Bottom left: GC content vs splice score. Bottom right: distributions of regional scores

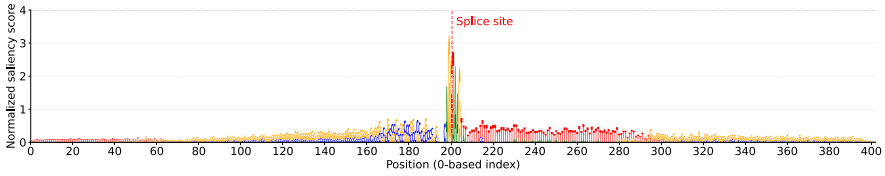

Fig. 10 Sequence logo for donor sequences containing single U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide GT shown at positions 200–201 using 0-based indexing.

Having multiple U2-type introns per sequence improves splice site prediction effectiveness

The second hypothesis dealt with the impact of the number of introns present within sequences. Our experimental results pointed to higher effectiveness in splice site prediction for datasets containing multiple U2-type introns per sequence. This observation is consistent with the underlying biological mechanisms that govern RNA splicing, where the presence of multiple introns within a gene sequence has been shown to positively impact splicing kinetics and gene expression regulation [14].

For donor splice sites, the sequence logos for single U2-type introns (Fig. 10) and for multiple U2-type introns (Fig. 11) show the GT dinucleotide at positions 200 and 201, which correspond to the donor splice site in the dataset. A notable difference emerges in the polypyrimidine tract distribution, where Fig. 11 shows prominent signals both upstream and downstream of the splice site, while Fig. 10 exhibits only one dominant polypyrimidine tract with high saliency scores in the downstream region of the splice site.

For acceptor splice sites, the sequence logos for single U2-type introns (Fig. 12) and for multiple U2-type introns (Fig. 13) show similar differences in polypyrimidine tract arrangement around the AG dinucleotide at positions 200 and 201. The dual presence

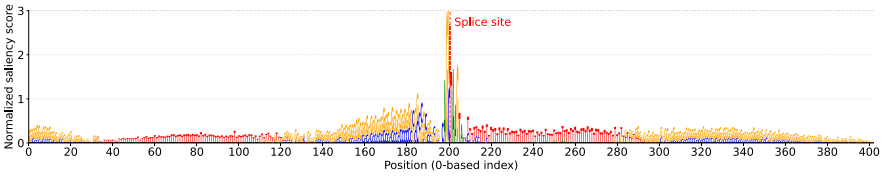

Fig. 11 Sequence logo for donor sequences containing multiple U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide GT shown at positions 200–201 using 0-based indexing.

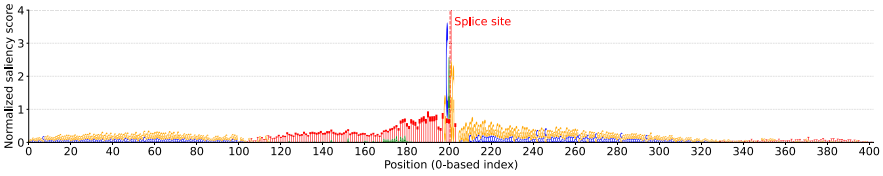

Fig. 12 Sequence logo for acceptor sequences containing single U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide AG shown at positions 200–201 using 0-based indexing.

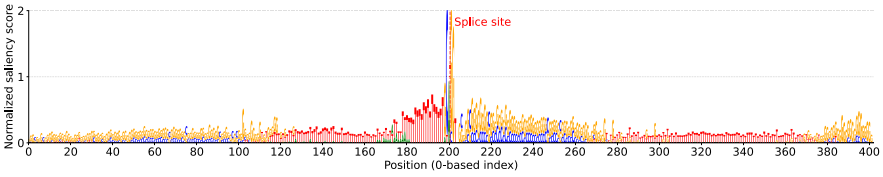

Fig. 13 Sequence logo for acceptor sequences containing multiple U2-type introns (nucleotide colors: A = green, C = blue, G = orange, T = red). Splice site dinucleotide AG shown at positions 200–201 using 0-based indexing.

of polypyrimidine tracts in sequences with multiple U2-type introns suggests a more sophisticated splicing regulation, facilitating the cooperative recruitment of splicing factors, particularly U2AF65, which is a key protein involved in pre-mRNA splicing [25]. From a computational perspective, these dual high-contribution regions provide additional discriminative features for splice site prediction, supporting our observed improvement in prediction effectiveness for sequences containing multiple U2-type introns.

Quantitative correlation analysis of single versus multiple U2-type intron showed that for donor sites (Fig. 14), the extended consensus showed strong correlations with splice site saliency in both categories (r=0.90, single; r=0.82, multiple; p<0.05). Upstream MAG correlations were moderate (r=0.16, single; r=0.39, multiple; p<0.05), whereas downstream RAGT correlations were weaker and negative in multiple U2-type intron sequences (r=0.14, single; r=−0.41, multiple; p<0.05). For acceptor sites (Fig. 15), branch point correlations remained modest (r=0.21, single; r=0.27, multiple; p<0.05), while polypyrimidine tract correlations were stronger (r=0.60, single; r=0.54, multiple; p<0.05). GC content correlations were weak across both categories (r=−0.10, single; r=−0.03, multiple), with the latter not reaching significance ($p\geq$0.05).

Prior research has indicated that early splicing events facilitate splicing of downstream introns on the same transcript [25]. Early splicing events possibly recruit necessary protein complexes, such as EJC, resulting in faster splicing of downstream introns. This mechanism suggests a synergistic effect, in which each splicing event incrementally improves the precision of subsequent splicing events within the same transcript.

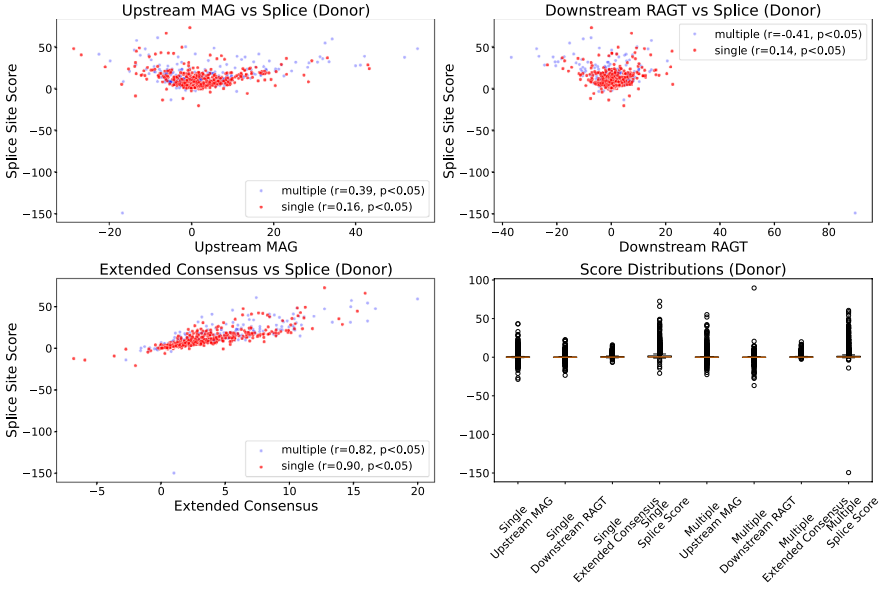

Fig. 14 Quantitative validation of model attention patterns for donor sites (single vs multiple). Colors indicate intron category: single (red) and multiple (blue). Each panel shows the relationship between regional normalized saliency scores and the splice-site score. The legend reports the Pearson correlation coefficient (r) and the two-sided *p*-value; $p<0.05$ denotes a statistically significant correlation at the 5% level, and $p\geq 0.05$ denotes a non-significant result. Top left: upstream `MAG` consensus vs splice score. Top right: downstream `RAGT` consensus vs splice score. Bottom left: extended donor consensus (`MAG GTRAGT`) vs splice score. Bottom right: distributions of regional scores

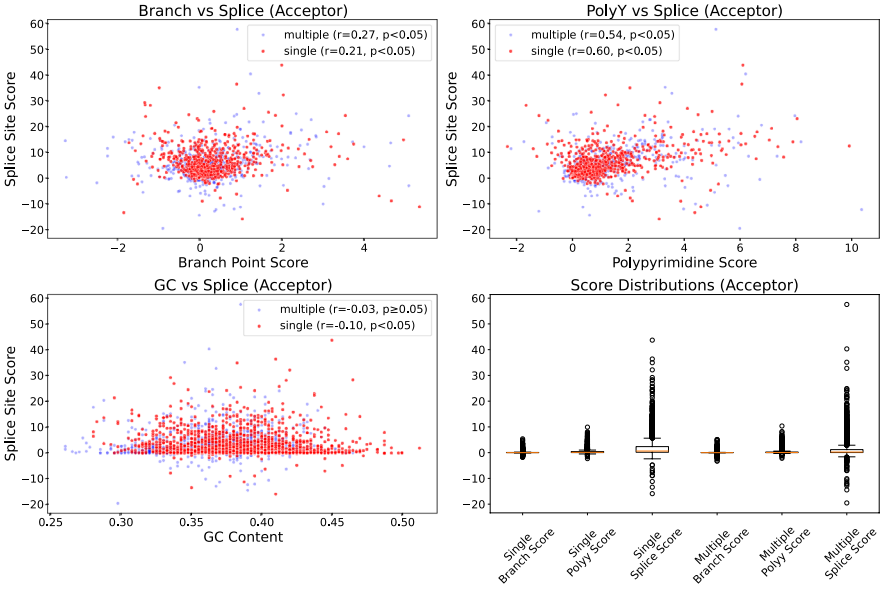

Fig. 15 Quantitative validation of model attention patterns for acceptor sites (single vs multiple). Colors indicate intron category: single (red) and multiple (blue). Each panel shows the relationship between regional normalized saliency scores and the splice-site score. The legend reports the Pearson correlation coefficient (*r*) and the two-sided *p*-value; $p<0.05$ denotes a statistically significant correlation at the 5% level, and $p\geq 0.05$ denotes a non-significant result. Top left: branch point adenine region vs splice score. Top right: polypyrimidine tract vs splice score. Bottom left: GC content vs splice score. Bottom right: distributions of regional scores

In addition, multiple introns can enhance the efficiency of co-transcriptional splicing (CTS) [56]. CTS refers to splicing that occurs during the transcription process. In *Arabidopsis thaliana*, individual loci such as *FLOWERING LOCUS C* (FLC) and DELAY OF GERMINATION 1 (DOG1) have shown CTS activities [17, 41]. The average CTS efficiency correlates strongly with the number of introns per gene and appears independent of gene length. Notably, RZ-1 proteins, which are RNA binding proteins, have been shown to promote CTS of genes with multiple introns at the chromatin level [56].

These studies demonstrate that, while no single intron may be crucial for expression, the cumulative presence of multiple introns can substantially enhance mRNA accumulation and gene expression in an additive way. This effect can also extend to the regulation of splicing factor abundance, indirectly influencing splice site selection accuracy [8].

The key observations and biological considerations are: (1) The models show higher predictive performance for sequences containing multiple U2-type introns compared to those containing a single intron. This may reflect that multi-intron sequences harbor a richer and more robust collection of splicing regulatory elements (such as multiple branch points and extended polypyrimidine tracts), which are critical for accurate splicing. (2) The difference in prediction performance is consistent with biological studies indicating that genes with multiple introns in *Arabidopsis thaliana* tend to display coordinated splicing regulation. For example, the conserved branch point-to-acceptor distance and the presence of several regulatory motifs are thought to contribute to efficient co-transcriptional splicing.

Downstream biological impact

Our findings on U2-type intron characteristics indicate potential avenues for practical application in important domains of plant biology.

For genome annotation in *Arabidopsis thaliana*, our findings provide modest improvements in splice site prediction accuracy. These improvements could be incorporated into existing annotation pipelines through context-dependent scoring adjustments, such as applying higher confidence weights to acceptor sites in short intron contexts or to splice sites in multi-intron gene models [22].

However, several important limitations must be acknowledged. First, the magnitude of improvement is modest and may not substantially impact overall annotation quality. Second, our analysis was limited to *Arabidopsis thaliana*, and direct transfer to other plant species requires validation, particularly for species with different median intron lengths or intron multiplicity distributions [52]. Third, implementation would require species-specific parameter optimization rather than direct application of our *Arabidopsis*-derived thresholds. The challenge of cross-species applicability is particularly important for practical implementation. Our *Arabidopsis*-based threshold of 90 bp for distinguishing short from long U2-type introns may not apply to other species. For example, rice (*Oryza sativa*) exhibits a median intron length of approximately 400 bp [38], while maize shows highly variable intron size distributions [26]. Similarly, the proportion of genes containing multiple introns varies significantly across plant lineages. Therefore, while our findings establish proof-of-concept evidence that intron characteristics influence splice site prediction accuracy, practical application to genome annotation tools requires species-specific validation studies to determine appropriate parameter settings for each target organism.

For genetic engineering and synthetic biology applications, our results provide computational insights that may inform intron design strategies, though they do not establish optimal intron configurations for biological function. Our findings indicate that U2-type short introns show improved acceptor site prediction accuracy and that sequences with multiple U2-type introns exhibit enhanced splice site predictability. However, computational predictability does not necessarily correlate with superior expression levels or splicing efficiency in engineered constructs [31, 43]. Therefore, while our analysis suggests certain intron features are associated with more robust splice site signals, experimental validation through transgenic studies would be required to determine their effectiveness for improving gene expression in synthetic biology applications [18].

Conclusions and directions for future research

This study investigated two key hypotheses regarding splice site prediction in *Arabidopsis thaliana*. First, we examined whether models trained on datasets with short U2-type introns would demonstrate higher effectiveness compared to those trained on datasets with long U2-type introns. Secondly, we tested whether datasets containing sequences with multiple U2-type introns would yield higher prediction effectiveness compared to those with single U2-type introns.

Our findings partially supported the first hypothesis, demonstrating that models trained on datasets containing short U2-type introns exhibited enhanced predictive capabilities for acceptor splice sites across all tested models. However, no substantial improvement was observed for donor splice sites, where performance remained consistent across intron length categories. The second hypothesis was fully supported, with the presence of multiple U2-type introns within sequences resulting in more effective splice site prediction across all models, particularly for acceptor sites.

Our computational analysis contributes to the field by providing systematic validation and quantification of known biological features affecting splice site recognition. While our study advances computational methods for splice site prediction, we emphasize that these findings represent computational predictability rather than direct biological mechanism understanding. The improved prediction accuracy for short introns and multi-intron sequences provides computational evidence supporting the functional importance of these features, but determining their causal roles in splicing efficiency requires experimental validation.

For practical applications, our findings provide computational baselines that may inform future studies, though several important limitations must be acknowledged. Our analysis was limited to *Arabidopsis thaliana*, and direct transfer to other plant species requires validation, particularly for species with different intron length distributions. While our computational insights could theoretically inform genome annotation pipeline modifications, the modest improvements we observed require careful evaluation of implementation complexity versus benefit. Similarly, while our results suggest certain intron features are associated with more predictable splice site patterns, experimental validation through transgenic studies would be required to determine their effectiveness for improving gene expression in synthetic biology applications.

Several methodological limitations should guide future work. Our saliency map analysis provided visual interpretation of sequence feature importance, but quantitative correlation analysis between saliency scores and prediction performance would strengthen

the biological relevance of our computational findings. Additionally, the relationship between computational prediction accuracy and actual splicing efficiency in biological systems remains to be experimentally determined. Future research should focus on validating these computational predictions across diverse plant species, developing quantitative frameworks for interpreting model attention mechanisms, and experimentally testing whether computationally robust splice sites translate to improved splicing efficiency in engineered constructs. Such work could bridge the gap between computational prediction and biological function, ultimately contributing to more effective tools for plant biotechnology and our understanding of transcriptome complexity. Additionally, given the limited understanding of alternative splicing regulation in plants, future work could explore how deep learning models can be adapted to identify and interpret alternative splicing events, ultimately contributing to a deeper understanding of transcriptome complexity.

Author Contributions
E.K. designed and implemented the project, and wrote the initial draft manuscript. S.J. contributed to the results analysis and literature review. S.Y. was responsible for the design of figures and also participated in the literature review. S.D. provided biological input and reviewed the manuscript. A.V.M. conducted the statistical review and reviewed the manuscript. W.D.N. contributed to the deep learning aspects and reviewed the manuscript.

Funding
Not applicable.

Data Availability
The annotated intron sequences can be downloaded at https://introndb.lerner.ccf.org/static/fasta/TAIR10_U2.fasta and the A. thaliana DNA sequences with splice sites can be downloaded from https://github.com/XueyanLiu-creator/DRANetSplicer/tree/main/data/dna_sequences. The models' implementation and the code used to conduct the experiments in this study are available at https://github.com/EspoirKabanga/Impact_of_U2type_Introns_on_Splice_Site_Prediction.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Conflict of interest
The authors declare no Conflict of interest.

Received: 16 July 2025 / Accepted: 29 October 2025
Published online: 28 November 2025

References

1. Akpokiro V, Oluwadare O, Kalita J. DeepSplicer: An Improved Method of Splice Sites Prediction using Deep Learning. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp 606–609, 2021. https://doi.org/10.1109/ICMLA52953.2021.00101
2. Akpokiro V, Martin T, Oluwadare O. EnsembleSplice: ensemble deep learning model for splice site prediction. BMC Bioinformatics. 2022;23(1):413. https://doi.org/10.1186/s12859-022-04971-w.
3. Albaradei S, Magana-Mora A, Thafar M, et al. Splice2Deep: An ensemble of deep convolutional neural networks for improved splice site prediction in genomic dna. Gene. 2020;763:100035. https://doi.org/10.1016/j.gene.2020.100035
4. Amit M, Donyo M, Hollander D, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. 2012. https://doi.org/10.1016/j.celrep.2012.03.013.
5. Angermueller C, Pärnamaa T, Parts L, et al. Deep learning for computational biology. Molecular Systems Biology. 2016;12(7):878. https://doi.org/10.15252/msb.20156651
6. Anil AT, Pandian R, Mishra SK. Introns with branchpoint-distant 3' splice sites: Splicing mechanism and regulatory roles. Biophys Chem. 2024;314:107307. https://doi.org/10.1016/j.bpc.2024.107307
7. Basu MK, Rogozin IB, Koonin EV. Primordial spliceosomal introns were probably u2-type. Trends Genet. 2008;24(11):525–8. https://doi.org/10.1016/j.tig.2008.09.002.
8. Bourdon V, Harvey A, Lonsdale DM. Introns and their positions affect the translational activity of mRNA in plant cells. EMBO Rep. 2001;2(5):394–8. https://doi.org/10.1093/embo-reports/kve090.
9. Chang N, Sun Q, Hu J, An C, Gao H. Large introns of 5 to 10 kilo base pairs can be spliced out in arabidopsis. Genes. 2017;8(8):200. https://doi.org/10.3390/genes8080200.

10. Chen W, Moore JM. The spliceosome: disorder and dynamics defined. Curr Opin Struct Biol. 2014;24:141–9. https://doi.org/10.1016/j.sbi.2014.01.009.
11. Chen ZJ, Wang J, Tian L, et al. The development of an arabidopsis model system for genome-wide analysis of polyploidy effects. Biol J Lin Soc. 2004;82(4):689–700. https://doi.org/10.1111/j.1095-8312.2004.00351.x.
12. Chung BYW, Simons C, Firth AE, et al. Effect of 5'UTR introns on gene expression in arabidopsis thaliana. BMC Genomics. 2006;7(1):120. https://doi.org/10.1186/1471-2164-7-120.
13. Clancy S. RNA splicing: Introns, Exons and Spliceosome. Nature. Education. 2008;1(1):31.
14. Crabb TL, Lam BJ, Hertel KJ. Retention of spliceosomal components along ligated exons ensures efficient removal of multiple introns. RNA. 2010;16(9):1786–96. https://doi.org/10.1261/rna.2186510.
15. Dewey CN, Rogozin IB, Koonin EV. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. BMC Genomics. 2006;7(1):311. https://doi.org/10.1186/1471-2164-7-311.
16. Dietrich RC, Incorvaia R, Padgett AR. Terminal intron dinucleotide sequences do not distinguish between u2- and u12-dependent introns. Mol Cell. 1997;1(1):151–60. https://doi.org/10.1016/S1097-2765(00)80016-7.
17. Dolata J, Guo Y, Kołowerzo A, et al. NTR 1 is required for transcription elongation checkpoints at alternative exons in arabidopsis. The EMBO J. 2015;34(4):544–58. https://doi.org/10.15252/embj.201489478
18. Gallegos JE, Rose AB. The enduring mystery of intron-mediated enhancement. Plant Sci. 2015;237:8–15. https://doi.org/10.1016/j.plantsci.2015.04.017.
19. Girardini KN, Olthof AM, Kanadia RN. Introns: the "dark matter" of the eukaryotic genome. Front Genet. 2023;14:1150212. https://doi.org/10.3389/fgene.2023.1150212.
20. Goodall GJ, Filipowicz W. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell. 1989;58(3):473–83. https://doi.org/10.1016/0092-8674(89)90428-5.
21. Hebsgaard SM, Korning PG, Tolstrup N, et al. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res. 1996;24(17):3439–52. https://doi.org/10.1093/nar/24.17.3439.
22. Hoff KJ, Lomsadze A, Borodovsky M, et al. Whole-Genome Annotation with BRAKER. In: Gene prediction: methods and protocols. Springer, 2019 p 65–95, https://doi.org/10.1007/978-1-4939-9173-0_5
23. Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol. 2006;23(12):2392–404. https://doi.org/10.1093/molbev/msl111.
24. James AB, Syed NH, Bordage S, et al. Alternative splicing mediates responses of the arabidopsis circadian clock to temperature changes. Plant Cell. 2012;24(3):961–81. https://doi.org/10.1105/tpc.111.093948.
25. Jia J, Long Y, Zhang H, et al. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. Nature Plants. 2020;6(7):780–8. https://doi.org/10.1038/s41477-020-0688-1.
26. Jiao Y, Peluso P, Shi J, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(7659):524–7. https://doi.org/10.1038/nature22971.
27. JiaYan W, JingFa X, LingPing W, et al. Systematic analysis of intron size and abundance parameters in diverse lineages. Science China Life Sciences. 2013;56(10):968–74. https://doi.org/10.1007/s11427-013-4540-y.
28. Karve R, Liu W, Willet SG, et al. The presence of multiple introns is essential for ERECTA expression in arabidopsis. RNA. 2011;17(10):1907–21. https://doi.org/10.1261/rna.2825811.
29. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint https://doi.org/10.48550/arXiv.1412.6980 (2004)
30. Lanchantin J, Singh R, Wang B, et al. Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks. Pac Symp Biocomput. 2017;22:254–65. https://doi.org/10.1142/9789813207813_0025.
31. Laxa M. Intron-mediated enhancement: a tool for heterologous gene expression in plants? Front Plant Sci. 2017;7:1977. https://doi.org/10.3389/fpls.2016.01977.
32. Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci. 2001;98(20):11193–8. https://doi.org/10.1073/pnas.201407298.
33. Liu X, Zhang H, Zeng Y, et al. DRANetSplicer: A Splice Site Prediction Model Based on Deep Residual Attention Networks. Genes. 2024;15(4):404. https://doi.org/10.3390/genes15040404
34. Marquez Y, Brown JW, Simpson C, et al. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. 2012;22(6):1184–95. https://doi.org/10.1101/gr.134106.111.
35. Martín G, Márquez Y, Mantica F, et al. Alternative splicing landscapes in arabidopsis thaliana across tissues and stress conditions highlight major functional differences with animals. Genome Biol. 2021;22(1):35. https://doi.org/10.1186/s13059-020-02258-y.
36. Moyer DC, Larue GE, Hershberger CE, et al. Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res. 2020;48(13):7066–78. https://doi.org/10.1093/nar/gkaa464.
37. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R. Intron retention is a major phenomenon in alternative splicing in Arabidopsis. The Plant Journal. 2004;39(6):877–85. https://doi.org/10.1111/j.1365-313X.2004.02172.x
38. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J. The TIGR rice genome annotation resource: improvements and new features. Nucleic acids research 35 2007 Jan 1;35(suppl_1):D883-7. https://doi.org/10.1093/nar/gkl976
39. Patel AA, Steitz JA. Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol. 2003;4(12):960–70. https://doi.org/10.1038/nrm1259.
40. Reddy AS, Marquez Y, Kalyna M, et al. Complexity of the alternative splicing landscape in plants. Plant Cell. 2013;25(10):3657–83. https://doi.org/10.1105/tpc.113.117523.
41. Rosa S, Duncan S, Dean C. Mutually exclusive sense-antisense transcription at FLC facilitates environmentally induced gene repression. Nat Commun. 2016;7(1):13031. https://doi.org/10.1038/ncomms13031.
42. Rose AB. Intron-mediated regulation of gene expression. Current Topics in Microbiology and Immunology. 2008;1:277–90. https://doi.org/10.1007/978-3-540-76776-3_15
43. Rose AB. Introns as gene regulators: a brick on the accelerator. Front Genet. 2019;9:672. https://doi.org/10.3389/fgene.2018.00672
44. Roy M, Kim N, Xing Y, et al. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA. 2008;14(11):2261–73. https://doi.org/10.1261/rna.1024908.

45. Sales-Lee J, Perry DS, Bowser BA, et al. Coupling of spliceosome complexity to intron diversity. Curr Biol. 2021;31(22):4898-4910.e4. https://doi.org/10.1016/j.cub.2021.09.004
46. Sharp PA, Burge CB. Classification of introns: U2-Type or U12-Type. Cell. 1997;91(7):875–9.
47. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations, 2014. https://doi.org/10.48550/arXiv.1312.6034
48. Szeghalmy S, Fazekas A. A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors. 2023;23(4):2333. https://doi.org/10.3390/s23042333
49. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in python. Bioinformatics. 2019;36(7):2272–4. https://doi.org/10.1093/bioinformatics/btz921
50. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. Nature. 2000;408(6814):796–815. https://doi.org/10.1038/35048692.
51. Tolstrup N, Rouzé P, Brunak S. A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res. 1997;25(15):3159–63. https://doi.org/10.1093/nar/25.15.3159.
52. Wang BB, Brendel V. Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci. 2006;103(18):7175–80. https://doi.org/10.1073/pnas.0602039103.
53. Wang R, Wang Z, Wang J, et al. SpliceFinder: ab initio prediction of splice sites using convolutional neural network. BMC Bioinformatics. 2019;20(23):652. https://doi.org/10.1186/s12859-019-3306-3.
54. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2–3):377–94. https://doi.org/10.1089/1066527041410418.
55. Zhang Q, Li H, Zhao XQ, et al. The evolution mechanism of intron length. Genomics. 2016;108(2):47–55. https://doi.org/10.1016/j.ygeno.2016.07.004.
56. Zhu D, Mao F, Tian Y, et al. The Features and Regulation of Co-transcriptional Splicing in Arabidopsis. Mol Plant. 2020;13(2):278–94. https://doi.org/10.1016/j.molp.2019.11.004.
57. Zuallaert J, Godin F, Kim M, et al. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. Bioinformatics. 2018;34(24):4180–8. https://doi.org/10.1093/bioinformatics/bty497.

Publisher's Note