Improving GAN-Generated Splice Site Sequences Through Conditional Frequency Guidance

Espoir Kabanga*
Center for Biosystems and Biotech Data Science
Ghent University Global Campus
Incheon, South Korea
and
IDLab, Department of Electronics and Information Systems
Ghent University
Ghent, Belgium
espoir.kabanga@ghent.ac.kr
*Corresponding author

Wesley De Neve
Center for Biosystems and Biotech Data Science
Ghent University Global Campus
Incheon, South Korea
and
IDLab, Department of Electronics and Information Systems
Ghent University
Ghent, Belgium
wesley.deneve@ghent.ac.kr

Arnout Van Messem
Department of Mathematics
Université de Liège
Liège, Belgium
arnout.vanmessem@uliege.be

***Abstract*—Generative adversarial networks (GANs) struggle to capture position-specific nucleotide constraints that define splice site identity. We present a frequency-blended framework that guides GAN synthesis by linearly combining model predictions with conditional empirical priors. We evaluate synthetic donor and acceptor splice sites from *Arabidopsis thaliana* and *Homo sapiens* through direct biological assessments (3-mer context and sequence logo) and functional validation using SpliceRover, a state-of-the-art splice site classifier. Our results show that frequency blending substantially improves biological fidelity and predictive performance compared to unguided generation, recovering position-specific motifs that GANs systematically miss. Data augmentation with 50% real and 50% frequency-blended synthetic sequences achieves baseline-level predictive performance, effectively halving real genomic data requirements while maintaining state-of-the-art classification accuracy.**

***Index Terms*—GAN, frequency blending, splice sites**

I. Introduction

In eukaryotes, the splicing machinery excises introns from precursor mRNA and ligates exons to produce functional transcripts [1], [2]. Splice boundaries, donor (5') and acceptor (3') sites, are represented by `GT` and `AG` dinucleotides, respectively [3], [4], and are flanked by species-specific conservation patterns that influence spliceosome recognition [5], [6]. Generation of artificial sequences with splice sites addresses multiple objectives: expanding training datasets [7], enabling privacy-preserving data sharing [8], benchmarking prediction algorithms [9], and facilitating comparative studies [10].

Classical approaches, including position weight matrices [11], Markov models [12], [13], and maximal dependence decomposition [14], capture local dependencies but have limited capacity to encode long-range relationships for functional splice sites [15]. Generative adversarial networks (GANs), which train a generator to produce sequences that fool a discriminator network through iterative competition [16], have successfully been applied to regulatory sequence generation [17], [18], promoters [19], and enhancers [20]. However, while the generator learns to produce plausible sequences and the discriminator evaluates their global realism, this adversarial framework does not explicitly enforce position-specific constraints observed in splice site recognition (conserved boundary dinucleotides, purine enrichment upstream of donors for U1 snRNP binding [21], and pyrimidine tracts upstream of acceptors for U2AF recognition [22]). We present a frequency blending strategy that incorporates position-dependent empirical priors into GAN outputs. By interpolating GAN predictions with positional nucleotide frequencies using a mixing weight λ, we modulate the sequence characteristics. This work includes: (1) a post-training blending mechanism that adjusts compositional properties while retaining learned patterns, (2) validation for *Arabidopsis thaliana* and *Homo sapiens* using compositional metrics and classification performance via SpliceRover [23], (3) (3) systematic evaluation of blending weights ($\lambda \in \{0.0, 0.25, 0.5, 0.75\}$) demonstrating that $\lambda = 0.5$ balances model predictions and empirical frequency priors, and (4) results indicating that the use of 50% real + 50% synthetic training data approximates baseline predictive performance while reducing real data requirements. Fig. 1 illustrates the workflow of the frequency-blended GAN framework. Random noise passes through the generator to produce sequence probabilities, which are blended with position-specific nucleotide frequencies extracted from real splice site sequences using the parameter value λ. The resulting synthetic sequences (with conserved splice site dinucleotides highlighted in red) are evaluated by the discriminator alongside real sequences. Adversarial training (dashed feedback loop) improves the generator performance over iterations.

II. Experimental Setup

A. Dataset and Preprocessing

We utilized splice site sequences from the DRANetSplicer dataset [24], comprising approximately $64,000$ sequences for *Arabidopsis thaliana* and $120,000$ sequences for *Homo*

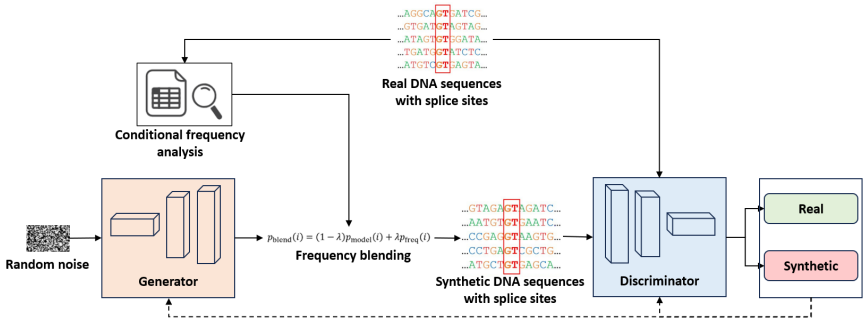

Fig. 1. Integration of frequency blending into a GAN-based approach for the generation of splice site sequences.

sapiens. Each sequence spans 402 base pairs, centered on annotated splice boundaries. The dataset provides positive and negative examples for both donor and acceptor sites, where positive sequences represent verified splice sites and negative sequences contain decoy GT/AG dinucleotides that do not function as true splice sites. For GAN training, we randomly sampled $50,000$ sequences from the *Arabidopsis thaliana* positive sets and $100,000$ sequences from the *Homo sapiens* positive sets, stratified by splice type (donor and acceptor). We reserved separate data for evaluation to prevent information leakage. All sequences were encoded as one-hot tensors with four channels representing nucleotides A, C, G, and T.

B. Position-Dependent Frequency Extraction

To guide the generator toward biologically plausible sequences, we extracted conditional nucleotide frequencies from training alignments. These frequencies quantify the probability of observing each nucleotide at a given position conditioned on its neighboring base. For a sequence $S = (s_0, s_1, \ldots, s_{L-1})$ of length $L = 402$ and with $s_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we computed: $P(s_i \mid s_{i-1})$, $0 < i \leq L-1$ and $P(s_i \mid s_{i+1})$, $0 \leq i < L-1$, representing the previous-neighbor and the next-neighbor conditioning, respectively. The previous-neighbor conditional probability is estimated as:

$$P(s_i = x \mid s_{i-1} = b) = \frac{\text{Count}(s_{i-1} = b, s_i = x)}{\text{Count}(s_{i-1} = b)}$$

where $b, x \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The next-neighbor probabilities are estimated analogously. These position-specific frequencies capture local sequence context including critical motifs such as purine enrichment upstream of donors and pyrimidine tracts upstream of acceptors.

C. GAN Architecture

Our GAN architecture follows the adversarial training framework [16], where a generator network maps 100-dimensional Gaussian noise vectors $\mathbf{z} \sim \mathcal{N}(0, I_{100})$ through fully connected layers to produce 402×4 sequence representations, while a discriminator network classifies sequences as real or synthetic. Both networks train alternately using the binary cross-entropy loss until the generator produces realistic sequences. To generate sequences, we pass noise through the trained generator to obtain logits, apply softmax normalization to produce position-wise nucleotide probabilities over {A, C, G, T}, and then blend these with empirical frequency priors. After blending, we enforce splice site dinucleotide constraints (GT for donors, AG for acceptors at positions 200–201) and sample discrete nucleotides independently at each position. Architecture specifications are provided in Table I and generator and discriminator losses can be seen in Fig. 2.

D. Frequency Blending Mechanism

Frequency blending operates exclusively during generation, leaving GAN training unchanged. After generating raw logits from the trained generator, we convert them to probabilities using softmax normalization. For each position $i \in$

TABLE I
GAN HYPERPARAMETERS AND ARCHITECTURE DETAILS.

Parameter	Value/Configuration
Generator	
Input dimension	100 (Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, I_{100})$)
Hidden layers	$64 \to 128 \to 256 \to 512$
Output dimension	1608 (reshaped to $(402, 4)$)
Activation function	LeakyReLU ($\alpha = 0.2$)
Normalization	Batch normalization
Dropout rate	0.3
Learning rate	5×10^{-5}
Discriminator	
Input dimension	1608 (flattened from $(402, 4)$)
Hidden layers	$512 \to 256 \to 128$
Output dimension	1 (binary classification)
Activation function	ReLU (hidden), Sigmoid (output)
Learning rate	2×10^{-4}
Training Configuration	
Loss function	Binary cross-entropy
Optimizer	Adam
Batch size	512
Number of epochs	50
Inference	
Noise sampling	100-dimensional standard Gaussian: $\mathbf{z} \sim \mathcal{N}(0, I_{100})$
Output processing	Softmax normalization + multinomial sampling

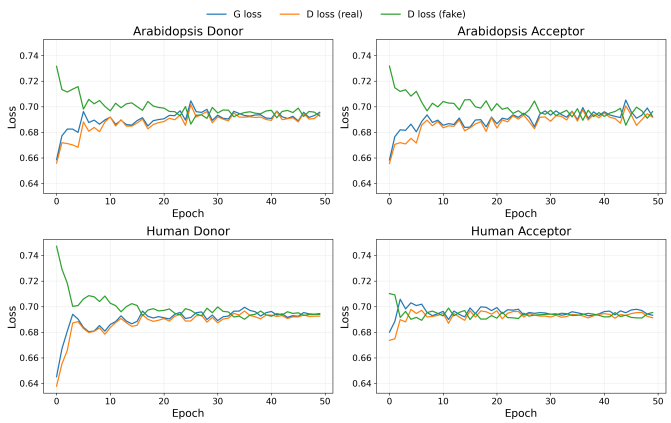

Fig. 2. Training loss curves of the GAN model across epochs for donor and acceptor splice site datasets from *Arabidopsis thaliana* and *Homo sapiens*. Shown are the generator loss (G loss), discriminator loss on real samples (D loss, real), and discriminator loss on synthetic samples (D loss, fake), illustrating stable adversarial training behavior across all settings.

$\{0, 1, \ldots, 401\}$, let $\mathbf{p}_{\text{model}}(i) \in \mathbb{R}^4$ denote the generator's predicted nucleotide probability distribution. To construct the conditional frequency prior $\mathbf{p}_{\text{freq}}(i) \in \mathbb{R}^4$, we use the most likely nucleotides from the generator's neighboring distributions: $s_{i-1} = \arg\max \mathbf{p}_{\text{model}}(i-1)$ and $si + 1 = \arg\max \mathbf{p}_{\text{model}}(i+1)$. We then query empirically-estimated conditional frequencies from training data, computing $\mathbf{p}_{\text{freq}}(i) = [P(s_i|s_{i-1}, i) + P(s_i|s_{i+1}, i)]/2$. We form the blended distribution as

$$\mathbf{p}_{\text{blend}}(i) = (1 - \lambda)\, \mathbf{p}_{\text{model}}(i) + \lambda\, \mathbf{p}_{\text{freq}}(i),$$

where $\lambda \in \{0.0, 0.25, 0.5, 0.75\}$ controls the balance between generator predictions and empirical frequencies. After blending all positions, nucleotides are sampled independently: $s_i \sim \text{Categorical}(\mathbf{p}_{\text{blend}}(i))$. We evaluated four configurations: $\lambda = 0.0$ (*No-Blend*), $\lambda = 0.25$ (model-dominated, 75% model), $\lambda = 0.5$ (equal weighting), and $\lambda = 0.75$ (frequency-dominated, 75% priors), with $\lambda = 0.5$ balancing positional constraints and generator-learned patterns.

After blending, we enforce splice site constraints by fixing positions 200–201 to `GT` (donors) or `AG` (acceptors), then sample discrete nucleotides independently at each position according to the categorical distribution defined by $\mathbf{p}_{\text{blend}}(i)$.

E. Evaluation Framework

We assessed synthetic sequence quality through two complementary strategies:

Direct Biological Evaluation. We computed compositional statistics across *Real* (authentic sequences) and synthetic sequences generated with varying blending weights ($\lambda \in \{0.0, 0.25, 0.5, 0.75\}$): (1) 3-mer frequencies in windows flanking splice junctions to evaluate local context preservation and (2) sequence logos visualizing position weight matrices around splice boundaries.

Functional Validation via Proxy Classification. We evaluated functional realism using the SpliceRover predictor [23]. We examined three scenarios with fixed training set size ($50,000$ sequences), using *Train-Real/Test-Real* (training and testing exclusively on real sequences) as the baseline:

Scenario 1 (*Train-Real/Test-Synthetic*): Train SpliceRover on real sequences, test on synthetic sequences. High accuracy indicates that synthetic sequences preserve discriminative features.

Scenario 2 (*Train-Synthetic/Test-Real*): Train SpliceRover on synthetic sequences, test on real sequences from the EnsembleSplice dataset [25]. This quantifies domain shift and whether blending improves generalization.

Scenario 3 (Data Augmentation): Train SpliceRover on mixtures of real and synthetic sequences ($\lambda = 0.5$) at ratios 10%, 25%, and 50% real (remainder synthetic), and test on real sequences from the EnsembleSplice dataset. This evaluates whether augmentation with synthetic sequences maintains predictive performance while reducing real data requirements.

SpliceRover is trained on balanced datasets with scenario-specific positives (real, synthetic, or mixtures) and real negatives from DRANet. GT/AG dinucleotides are enforced during generation, and conditional frequency priors are computed from the full GAN training set (50,000 *Arabidopsis thaliana*, 100,000 *Homo sapiens*), not from downstream training subsets. We report four metrics: Accuracy (proportion correct), F1-score (harmonic mean of precision and recall), MCC (correlation between predictions and labels), and AUROC (discriminative performance across thresholds).

III. Results and Discussion

A. Direct Evaluation

We assessed biological fidelity through 3-mer context analysis and sequence logo visualization across $\lambda \in \{0.0, 0.25, 0.5, 0.75\}$. Functional splice sites exhibit characteristic patterns: purine-rich regions upstream of donors facilitate U1 snRNP recognition [21], while pyrimidine-rich tracts upstream of acceptors enable U2AF binding [22].

3-mer frequencies reveal progressive improvement with increasing λ (Fig. 3). In *Arabidopsis thaliana* donors, *Real* sequences show enrichment of purine-rich 3-mers (`AAG`, `CAG`) upstream of splice junctions. Unblended generation ($\lambda = 0.0$) substantially underrepresents these motifs, with progressive recovery at $\lambda = 0.25$, $\lambda = 0.5$, and strongest approximation at $\lambda = 0.75$. Acceptor sites show similar trends for `T`-rich upstream contexts (`TGC`). In *Homo sapiens*, frequency blending progressively improves motif recovery for both splice types.

Sequence logos confirm these patterns across all λ values (Fig. 4). All conditions preserve canonical dinucleotides (`GT` for donors, `AG` for acceptors) at positions 200–201. In *Arabidopsis thaliana* donors, *Real* sequences display strong upstream purine enrichment and downstream `A`-bias. These features are substantially attenuated at $\lambda = 0.0$, partially recovered at $\lambda = 0.25$ and $\lambda = 0.5$, and most closely approximate *Real* patterns at $\lambda = 0.75$. For acceptors, increasing λ progressively recovers upstream `T`-enrichment. In *Homo sapiens*, $\lambda \geq 0.5$ effectively recovers polypyrimidine tracts upstream of acceptors and purine-rich contexts for donors.

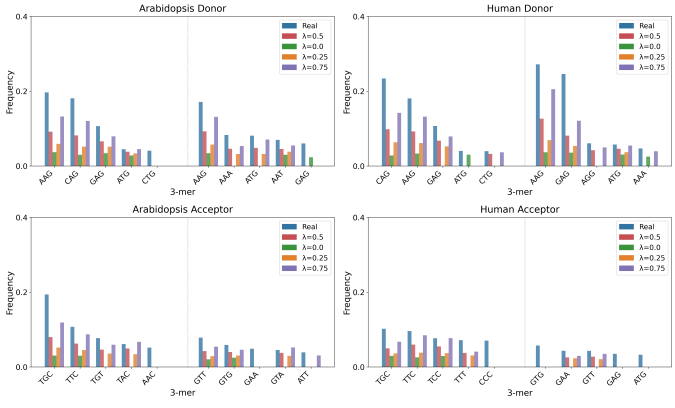

Fig. 3. 3-mer frequencies at positions 197 (upstream) and 202 (downstream) of splice junctions. Blue color represents *Real* sequences, red represents sequences with $\lambda = 0.5$, green *No-Blend* sequences with $\lambda = 0$, orange represents sequences generated with blending weight $\lambda = 0.25$, and purple for sequences generated with blending weight $\lambda = 0.75$. *Arabidopsis thaliana* (left) and *Homo sapiens* (right) for donor (top) and acceptor (bottom) splice sites.

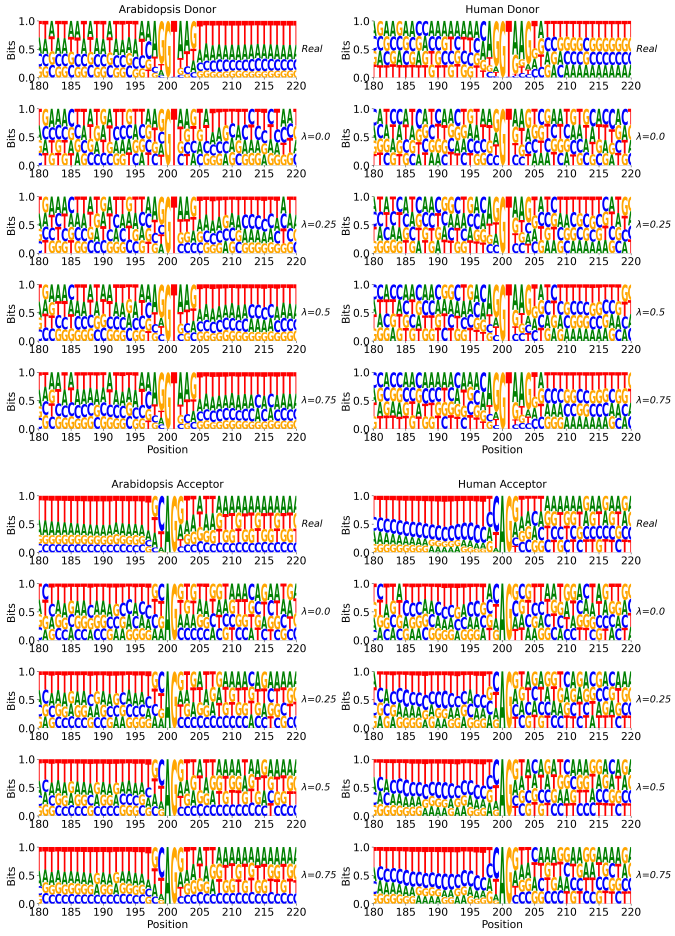

Fig. 4. Sequence logos of nucleotide conservation around splice junctions (positions 180–220; junction at 200–201). Left: *Arabidopsis thaliana* donors (top) and acceptors (bottom). Right: *Homo sapiens* donors (top) and acceptors (bottom). From top to bottom: *Real*, *No-Blend* ($\lambda = 0.0$), $\lambda = 0.25$, $\lambda = 0.5$, and $\lambda = 0.75$. Letter heights indicate nucleotide frequencies (bits).

These results demonstrate that unguided generation ($\lambda = 0.0$) captures global sequence properties but misses position-specific nucleotide constraints. Frequency blending progressively incorporates positional information, with higher λ values providing stronger enforcement of empirical motifs while $\lambda = 0.5$ balances fidelity and diversity.

B. Indirect Evaluation

In Scenario 1 (*Train-Real/Test-Synthetic*), we trained SpliceRover on real sequences and tested on synthetic sequences generated across $\lambda \in \{0.0, 0.25, 0.5, 0.75\}$ (Table II).

SpliceRover performance improves progressively with increasing λ. When tested on sequences generated at $\lambda = 0.0$, SpliceRover achieves near-random classification performance across species and splice types (Accuracy: 0.4925–0.5466, F1-scores: 0.0538–0.2143, MCC: -0.0401–0.1749, AUROC: 0.6036–0.7979), indicating that these sequences fail to preserve discriminative features. At $\lambda = 0.5$, performance improves substantially (Accuracy: 0.6974–0.7832, F1-scores: 0.5842–0.7414, MCC: 0.4706–0.5985, AUROC: 0.8416–0.9165), though remaining below baseline (*Train-Real/Test-Real*). Testing on sequences generated at $\lambda = 0.75$ yields the strongest performance (Accuracy: 0.7788–0.8631, F1-scores: 0.7266–0.8509, MCC: 0.6032–0.7360, AUROC: 0.9042–0.9536), approaching but not reaching baseline levels (Accuracy: 0.9594, F1-score: 0.9595, MCC: 0.9188, AUROC: 0.9899).

The highest performance occurs for *Homo sapiens* donors at $\lambda = 0.75$ (Accuracy: 0.8631, F1-score: 0.8509, MCC: 0.7360, AUROC: 0.9536). However, SpliceRover performance on all synthetic sequences remains below baseline, indicating incomplete capture of discriminative features present in real sequences.

In Scenario 2 (*Train-Synthetic/Test-Real*), we trained SpliceRover on synthetic sequences and tested on real sequences to quantify domain shift between synthetic and real sequence distributions (Table III).

For *Arabidopsis thaliana*, performance improves progressively with increasing λ but remains far below baseline (*Train-Real/Test-Real*). When trained on sequences generated at $\lambda = 0.0$, SpliceRover achieves near-random performance (Donors - Accuracy: 0.4962, F1-score: 0.0329, MCC: -0.0266, AUROC: 0.6437; Acceptors - Accuracy: 0.4948, F1-score: 0.0081, MCC: -0.0545, AUROC: 0.6421). Training on sequences generated at $\lambda = 0.5$ shows modest improvement (Donors - Accuracy: 0.5529, F1-score: 0.2296, MCC: 0.1948, AUROC: 0.7602; Acceptors - Accuracy: 0.5481, F1-score: 0.2093, MCC: 0.1865, AUROC: 0.7850). The strongest transfer occurs at $\lambda = 0.75$ (Donors - Accuracy: 0.6733, F1-score: 0.5433, MCC: 0.4216, AUROC: 0.8618; Acceptors - Accuracy: 0.5768, F1-score: 0.2948, MCC: 0.2560, AUROC: 0.8154), though remaining substantially below baseline (Accuracy: 0.9569, F1-score: 0.9573, MCC: 0.9140, AUROC: 0.9904).

For *Homo sapiens*, domain shift proves severe across all λ values. SpliceRover achieves near-random performance re-

gardless of blending weight (Donors - Accuracy: 0.5039–0.5111, F1-score: 0.0234–0.0598, MCC: 0.0435–0.0789; Acceptors - Accuracy: 0.5016–0.5028, F1-score: 0.0092–0.0156, MCC: 0.0231–0.0396). Notably, AUROC values for donors at higher λ decrease below $\lambda = 0.0$ (from 0.6092 to 0.5484), approaching random classification.

These results demonstrate fundamental limits to using synthetic sequences as complete replacements for real genomic data, with persistent performance gaps indicating systematic under-representation of sequence diversity and complex feature interactions.

In Scenario 3 (Data Augmentation), we evaluated whether augmenting limited real sequences with frequency-blended synthetic sequences maintains predictive performance. We trained SpliceRover on fixed-size datasets ($50,000$ sequences) containing varying proportions of real sequences (10%, 25%, 50%), with the remainder composed of synthetic sequences generated at $\lambda = 0.5$, and then tested on real sequences (Table IV).

Training on 50% real sequences augmented with 50% synthetic sequences achieves near-baseline performance across both species and splice types. For *Arabidopsis thaliana* donors, this configuration reaches Accuracy: 0.9581, F1-score: 0.9577, MCC: 0.9164, AUROC: 0.9872, closely matching baseline (Accuracy: 0.9594, F1-score: 0.9595, MCC: 0.9188, AUROC: 0.9899). Acceptor sites show similar results (Accuracy: 0.9425, F1-score: 0.9422, MCC: 0.8850, AUROC: 0.9810). For *Homo sapiens*, 50% real augmented with 50% synthetic exceeds baseline for donors (Accuracy: 0.9601 vs. 0.9594, F1-score: 0.9603 vs. 0.9600, MCC: 0.9202 vs. 0.9191, AUROC: 0.9899 vs. 0.9916) and approximates baseline for acceptors (Accuracy: 0.9514 vs. 0.9595, F1-score: 0.9502 vs. 0.9591, MCC: 0.9040 vs. 0.9192, AUROC: 0.9870 vs. 0.9890).

Performance degrades progressively as real fraction decreases. At 25% real augmented with 75% synthetic: *Arabidopsis thaliana* donors (Accuracy: 0.9523, F1-score: 0.9518, MCC: 0.9048, AUROC: 0.9839), acceptors (Accuracy: 0.9316,

TABLE II
SCENARIO 1 (*Train-Real/Test-Synthetic*): SPLICEROVER PERFORMANCE WHEN TRAINED ON REAL SEQUENCES AND TESTED ON GAN-GENERATED SEQUENCES. BASELINE REPRESENTS *Train-Real/Test-Real*.

Metric	**Baseline**	$\lambda = 0.0$ *(No-Blend)*	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
Arabidopsis thaliana – **Donor**					
Accuracy	0.9594	0.4925	0.6015	0.7058	0.8016
F1 Score	0.9595	0.0538	0.3825	0.6075	0.7653
MCC	0.9188	-0.0401	0.2880	0.4754	0.6342
AUROC	0.9899	0.6036	0.7659	0.8416	0.9042
Arabidopsis thaliana – **Acceptor**					
Accuracy	0.9569	0.5069	0.6188	0.7170	0.8129
F1 Score	0.9573	0.1206	0.4331	0.6328	0.7842
MCC	0.9140	0.0291	0.3143	0.4884	0.6494
AUROC	0.9904	0.6829	0.7980	0.8604	0.9173
Homo sapiens – **Donor**					
Accuracy	0.9594	0.5249	0.7106	0.7832	0.8631
F1 Score	0.9600	0.1808	0.6220	0.7414	0.8509
MCC	0.9191	0.0917	0.4767	0.5985	0.7360
AUROC	0.9916	0.7668	0.8779	0.9165	0.9536
Homo sapiens – **Acceptor**					
Accuracy	0.9595	0.5466	0.6303	0.6974	0.7788
F1 Score	0.9591	0.2143	0.4403	0.5842	0.7266
MCC	0.9192	0.1749	0.3547	0.4706	0.6032
AUROC	0.9890	0.7979	0.8629	0.8955	0.9308

TABLE III
SCENARIO 2 (*Train-Synthetic/Test-Real*): SPLICEROVER PERFORMANCE WHEN TRAINED ON GAN-GENERATED SEQUENCES AND TESTED ON REAL SEQUENCES. BASELINE REPRESENTS *Train-Real/Test-Real*.

Metric	**Baseline**	$\lambda = 0.0$ *(No-Blend)*	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
Arabidopsis thaliana – **Donor**					
Accuracy	0.9594	0.4962	0.5009	0.5529	0.6733
F1 Score	0.9595	0.0329	0.0443	0.2296	0.5433
MCC	0.9188	-0.0266	0.0059	0.1948	0.4216
AUROC	0.9899	0.6437	0.6569	0.7602	0.8618
Arabidopsis thaliana – **Acceptor**					
Accuracy	0.9569	0.4948	0.4974	0.5481	0.5768
F1 Score	0.9573	0.0081	0.0120	0.2093	0.2948
MCC	0.9140	-0.0545	-0.0276	0.1865	0.2560
AUROC	0.9904	0.6421	0.6808	0.7850	0.8154
Homo sapiens – **Donor**					
Accuracy	0.9594	0.5039	0.5111	0.5069	0.5051
F1 Score	0.9600	0.0234	0.0598	0.0411	0.0289
MCC	0.9191	0.0435	0.0789	0.0580	0.0525
AUROC	0.9916	0.6092	0.6336	0.5855	0.5484
Homo sapiens – **Acceptor**					
Accuracy	0.9595	0.5018	0.5023	0.5016	0.5028
F1 Score	0.9591	0.0148	0.0144	0.0092	0.0156
MCC	0.9192	0.0231	0.0330	0.0283	0.0396
AUROC	0.9890	0.5370	0.5343	0.5044	0.6219

TABLE IV
SCENARIO 3 (DATA AUGMENTATION): SPLICEROVER PERFORMANCE WHEN TRAINED ON MIXTURES OF REAL AND GAN-GENERATED SEQUENCES ($\lambda = 0.5$), TESTED ON REAL SEQUENCES. BASELINE REPRESENTS TRAINING ON 100% REAL SEQUENCES.

Metric	**Baseline (100% Real)**	10% Real	25% Real	50% Real
Arabidopsis thaliana – **Donor**				
Accuracy	0.9594	0.9279	0.9523	0.9581
F1 Score	0.9595	0.9246	0.9518	0.9577
MCC	0.9188	0.8590	0.9048	0.9164
AUROC	0.9899	0.9779	0.9839	0.9872
Arabidopsis thaliana – **Acceptor**				
Accuracy	0.9569	0.9027	0.9316	0.9425
F1 Score	0.9573	0.8957	0.9290	0.9422
MCC	0.9140	0.8127	0.8656	0.8850
AUROC	0.9904	0.9719	0.9796	0.9810
Homo sapiens – **Donor**				
Accuracy	0.9594	0.9031	0.9361	0.9601
F1 Score	0.9600	0.8939	0.9329	0.9603
MCC	0.9191	0.8187	0.8760	0.9202
AUROC	0.9916	0.9863	0.9879	0.9899
Homo sapiens – **Acceptor**				
Accuracy	0.9595	0.9047	0.8822	0.9514
F1 Score	0.9591	0.8958	0.8677	0.9502
MCC	0.9192	0.8216	0.7835	0.9040
AUROC	0.9890	0.9838	0.9848	0.9870

F1-score: 0.9290, MCC: 0.8656, AUROC: 0.9796); *Homo sapiens* donors (Accuracy: 0.9361, F1-score: 0.9329, MCC: 0.8760, AUROC: 0.9879), acceptors (Accuracy: 0.8822, F1-score: 0.8677, MCC: 0.7835, AUROC: 0.9848). At 10% real augmented with 90% synthetic, gaps widen further, particularly for *Homo sapiens* acceptors (Accuracy: 0.9047, F1-score: 0.8958, MCC: 0.8216, AUROC: 0.9838), though AUROC values remain high (> 0.97) across all conditions.

The aforementioned results demonstrate that frequency-blended synthetic sequences serve effectively as data augmentation when combined with adequate real sequences. Training with 50% real and 50% synthetic sequences maintains state-of-the-art predictive performance while halving real genomic data requirements. However, minimal real sequences (10%) prove insufficient to anchor learning, consistent with the findings of Scenario 2 that synthetic sequences alone cannot fully capture the complexity of real sequences.

IV. Conclusion and Future Work

We presented a frequency-blended framework that guides GAN-based splice site generation through conditional empirical priors. Systematic evaluation across blending weights ($\lambda \in \{0.0, 0.25, 0.5, 0.75\}$) demonstrates progressive improvement in biological fidelity and predictive performance, with $\lambda = 0.5$ balancing empirical constraints and generator-learned patterns. Frequency blending recovers position-specific motifs critical for splice recognition that unguided generation systematically misses.

Data augmentation with 50% real and 50% synthetic sequences achieves baseline-level performance, halving real genomic data requirements while maintaining state-of-the-art predictive accuracy. However, Scenario 2 reveals fundamental transfer limitations: *Homo sapiens* shows minimal improvement even at high λ values, while *Arabidopsis thaliana* exhibits progressive gains. This species-specific difference likely reflects greater sequence complexity in mammalian splice sites, where variable-length polypyrimidine tracts, branch point positioning, and distal regulatory elements introduce long-range dependencies that local conditional frequencies cannot adequately capture. Acceptor sites prove particularly challenging across both species, consistent with their greater biological complexity beyond immediately flanking positions.

Future work will extend frequency blending to other generative architectures (VAEs, diffusion models) and compare against traditional methods (position weight matrices, Markov models). Architectural improvements to the GAN itself, including convolutional or attention-based designs that explicitly model long-range dependencies, may better capture complex splice site features. Additional evaluation metrics (GC content, nucleotide conservation profiles, additional splice site predictors) and experimental validation through in vitro splicing assays would provide comprehensive assessment of functional realism.

References

[1] P. A. Sharp, "The discovery of split genes and RNA splicing," Trends in Biochemical Sciences, vol. 30, no. 6, pp. 279–281, 2005.

[2] M. C. Wahl, C. L. Will, and R. Lührmann, "The Spliceosome: Design Principles of a Dynamic RNP Machine," Cell, vol. 136, no. 4, pp. 701–718, 2009.

[3] S. M. Mount, "A catalogue of splice junction sequences," Nucleic Acids Research, vol. 10, no. 2, pp. 459–472, 1982.

[4] M. Burset, I. A. Seledtsov, and V. V. Solovyev, "Analysis of canonical and non-canonical splice sites in mammalian genomes," Nucleic Acids Research, vol. 28, no. 21, pp. 4364–4375, 2000.

[5] A. Churbanov, I. B. Rogozin, J. S. Deogun, and H. Ali, "Method of predicting Splice Sites based on signal interactions," Biology Direct, vol. 1, no. 1, p. 10, 2006.

[6] K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz et al., "Predicting Splicing from Primary Sequence with Deep Learning," Cell, vol. 176, no. 3, pp. 535–548, 2019.

[7] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in International Conference on Machine Learning, 2017, pp. 3145–3153.

[8] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," Circulation: Cardiovascular Quality and Outcomes, vol. 12, no. 7, p. e005122, 2019.

[9] G. Yeo and C. B. Burge, "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals," in Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, 2003, pp. 322–331.

[10] I. Korf, "Gene finding in novel genomes," BMC Bioinformatics, vol. 5, no. 1, p. 59, 2004.

[11] G. D. Stormo, "DNA binding sites: representation and discovery," Bioinformatics, vol. 16, no. 1, pp. 16–23, 2000.

[12] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models," Nucleic Acids Research, vol. 26, no. 2, pp. 544–548, 1998.

[13] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding Genes in DNA with a Hidden Markov Model," Journal of Computational Biology, vol. 4, no. 2, pp. 127–141, 1997.

[14] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," Journal of Molecular Biology, vol. 268, no. 1, pp. 78–94, 1997.

[15] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey, "Deciphering the splicing code," Nature, vol. 465, no. 7294, pp. 53–59, 2010.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in Advances in Neural Information Processing Systems, vol. 27, 2014.

[17] N. Killoran, L. J. Lee, A. Delong, D. Duvenaud, and B. J. Frey, "Generating and designing DNA with deep generative models," arXiv preprint arXiv:1712.06148, 2017.

[18] A. Gupta and J. Zou, "Feedback GAN for DNA optimizes protein functions," Nature Machine Intelligence, vol. 1, no. 2, pp. 105–111, 2019.

[19] Y. Wang, H. Wang, L. Wei, S. Li, L. Liu, and X. Wang, "Synthetic promoter design in *Escherichia coli* based on a deep generative network," Nucleic Acids Research, vol. 48, no. 12, pp. 6403–6412, 2020.

[20] C. Yin, S. Castillo-Hair, G. W. Byeon, P. Bromley, W. Meuleman, and G. Seelig, "Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity," Cell Systems, vol. 16, no. 7, 2025.

[21] C. B. Burge, "Splicing of precursors to mRNAs by the spliceosomes," in The RNA World, 1999, pp. 525–560.

[22] P. D. Zamore and M. R. Green, "Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor," Proceedings of the National Academy of Sciences, vol. 86, no. 23, pp. 9243–9247, 1989.

[23] J. Zuallaert, F. Godin, M. Kim, A. Soete, Y. Saeys, and W. De Neve, "SpliceRover: interpretable convolutional neural networks for improved splice site prediction," Bioinformatics, vol. 34, no. 24, pp. 4180–4188, 2018.

[24] X. Liu, H. Zhang, Y. Zeng, X. Zhu, L. Zhu, and J. Fu, "DRANetSplicer: A Splice Site Prediction Model Based on Deep Residual Attention Networks," Genes, vol. 15, no. 4, p. 404, 2024.

[25] I. Akpokiro and O. Oluwadaré, "Ensemblesplice: ensemble deep learning model for splice site prediction," BMC Bioinformatics, vol. 23, no. 1, pp. 1–14, 2022.