BENCHMARKING FINE-TUNED RNA LANGUAGE MODELS FOR INTRONIC BRANCH POINT PREDICTION

Pablo Rodenas Ruiz[1,*] **Ali Saadat**[1,*] **Timothy T. Tran**[2,1,*] **Oliver Müller Smedt**[1,*]
Peng Zhang[3] **Jacques Fellay**[1]

[1]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]University of Washington, Seattle, WA, United States
[3]The Rockefeller University, New York, NY, United States.
jacques.fellay@epfl.ch

ABSTRACT

Accurate prediction of RNA branch points is critical for understanding splicing mechanisms and identifying variants that may lead to genetic diseases. Despite their biological importance, few computational methods have been developed for reliably identifying branch points. In this work, we fine-tune several RNA language models for branch point prediction. The top-performing model, ERNIE-RNA, achieved an F_1 score of 0.811, a sequence accuracy of 0.790, and an average precision score of 0.868, outperforming previous leading models. These results showcase the potential of RNA-specific language models in capturing the subtle sequence features relevant to splicing. Our findings suggest that extended training and hyperparameter tuning could yield additional performance gains, positioning this study as a strong baseline for future research in RNA splicing.

1 INTRODUCTION

Branch points (BPs) are short motifs, usually a single nucleotide, within non-coding sequences (introns) of pre-mRNA, which play a crucial role in the RNA splicing mechanism. During splicing, the spliceosome removes introns and ligates coding sequences (exons), forming mature mRNA (Clancy, 2008). The key splicing motifs involved in this recognition are the 5' splice site (donor site), the 3' splice site (acceptor site), and the BP site (De Conti et al., 2013), as depicted in Fig. 1. The intronic sequence is cut from the RNA and bent on itself at the BP forming a lariat. The remaining RNA of the exonic sequences is then ready to be translated into proteins.

Figure 1: RNA splicing process and key splicing motifs.

Genetic variants that alter splicing motifs, including BPs, can prevent proper splicing (Douglas & Wood, 2011). Incorrect splicing can result in modified mRNA sequences, affecting the final proteins synthesized in the translation process. Even minor alterations in a protein's amino acid sequence can lead to genetic diseases. While most of the 5' and 3' splice site sequences have been experimentally mapped, BP sites exhibit considerable variation between different species and even between introns of the same organism (Xie et al., 2023). Due to the high variability in their positions, accurately predicting BPs and interpreting changes in their positions remain a critical challenge for molecular diagnosis and genetic disease research.

In this study, we fine-tuned various RNA language models on a dataset of BP positions and measured their performance in predicting BPs. Accurate BP prediction can support a range of downstream applications, such as assessing the impact of intronic variants that may lead to genetic diseases.

[*]Equal contribution.

When integrated with existing computational tools and other databases (Saadat & Fellay, 2024; 2025), these models can contribute to a more comprehensive understanding of how genetic variants influence disease mechanisms.

2 DATASET

This work utilizes an experimentally validated subset of the curated dataset from BPHunter (Zhang et al., 2022), consisting of more than 170,000 intronic DNA sequences with a single BP per sequence, derived from regions transcribed into pre-mRNA. Each sequence consists of NTs represented as A, G, T, or C, with a median sequence length of 1,558 NTs. The distribution of sequence lengths can be seen in Appendix A. BP annotations are provided as a binary sequence, where the position corresponding to the BP is labeled with 1 and all other positions with 0.

The dataset was divided into train, validation, and test sets according to the chromosome number. Specifically, chromosomes 9 and 10 were assigned to the validation set, chromosomes 8 and 11 to the test set, and the remaining chromosomes to the train set, resulting in a train/validation/test split of 83%, 8%, and 9%, respectively. This chromosome-based split prevents any overlap between the sequences in the training and evaluation sets, a strategy also used during the training of SpliceBERT (Chen et al., 2024).

Analysis of the training data highlighted clear trends in the position of the BPs. First, most BPs are located 15 to 40 NTs from the 3' splice site, aligning with previously known BP motifs. Secondly, adenine (A) is the nitrogenous base at most BP sites (95%). This enabled the creation of a baseline model that makes naive BP predictions, referred to as the "baseline" (see Appendix A).

3 MODELS

To fine-tune several models, we used MultiMolecule (Chen & Zhu, 2024), a library that provides a comprehensive collection of foundational RNA language models and tokenizers. Among the models we selected from this library, all have been pre-trained on intronic sequences (except UTR-LM), and contain only one NT per token, making them well-suited for BP detection. These models employ a BERT-style architecture, which is particularly effective for RNA sequence modeling because it effectively captures complex long-range dependencies and contextual relationships within NT sequences (Akiyama & Sakakibara, 2022). By leveraging patterns learned during pre-training, these models can be easily adapted for downstream tasks, such as identifying BPs. An overview of the key parameters and information for each model is provided in Table 1. The details of the fine-tuning procedure are provided in Appendix B.

Table 1: Key parameters of the RNA language models. All datasets are not limited to human sequences unless otherwise specified.

Model	**Size**	**Seq. Len.**	**Dataset Information**
RNABERT Akiyama & Sakakibara (2022)	0.48M	440	Human ncRNA (RNAcentral), SSL
UTR-LM (Chu et al., 2023)	1.21M	1022	5' UTR, SSL and SL
SpliceBERT (Chen et al., 2024)	19.72M	512	Vertebrate pre-mRNA (UCSC Genome Browser)
ERNIE-RNA (Yin et al., 2024)	85.67M	1024	ncRNA (RNAcentral), SSL
RNA-MSM (Zhang et al., 2023)	95.92M	1024	ncRNA and CREs (Rfam)
RNA-FM (Shen et al., 2024)	99.52M	1024	ncRNA (RNAcentral), SSL

4 RESULTS AND DISCUSSION

We evaluated model performance using the F_1 score, average precision (AP), Matthew's correlation coefficient (MCC), and the fraction of sequences where all tokens were correctly labeled (Seq. acc.). The performance metrics of the trained models on the test set are reported in Table 2 and the precision-recall curves are shown in Fig. 2. All fine-tuned models outperform the baseline, indicating that they are learning more complex relationships between the intron sequence and BP

Table 2: Performance metrics measured on the test set. The best performing model is showed in bold and the models are ordered by the number of parameters.

Model	Loss	F_1	Seq. Acc.	AP	MCC
Baseline	0.694	0.420	0.177	0.308	0.428
RNABERT	0.0066	0.555	0.412	0.556	0.558
UTR-LM	0.0042	0.585	0.414	0.541	0.597
SpliceBERT	0.0025	0.803	0.761	0.869	0.802
ERNIE-RNA	**0.0014**	**0.811**	**0.790**	**0.868**	**0.811**
RNA-MSM	0.0043	0.580	0.420	0.548	0.594
RNA-FM	0.0043	0.602	0.434	0.551	0.615

position. The best-performing model is ERNIE-RNA, followed closely by SpliceBERT, which both significantly outperform the others. The sequence accuracy during training is detailed in Appendix C, showing a clear trend of improvement with more training steps.

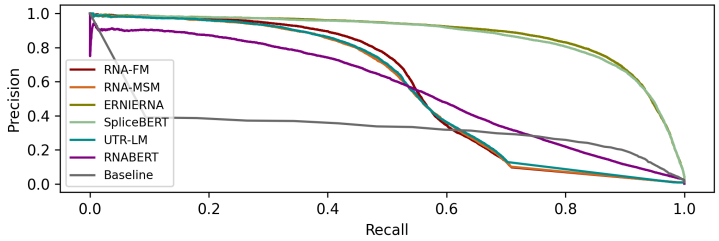

Figure 2: Precision-recall curves of all models, evaluated on the test set.

The fine-tuned models demonstrated strong performance, with ERNIE-RNA being the top performer, achieving an F_1 score of 0.811 and a sequence accuracy of 0.790. Moreover, the fine-tuned ERNIE-RNA and SpliceBERT models achieved an AP score of 0.868 and 0.869 respectively, outperforming the original SpliceBERT and SpliceBERT-human (Chen et al., 2024) by around 12.5% and 14.6%, respectively. This improvement could be due to a higher quality and more diverse dataset. However, it should be noted that the test set used for SpliceBERT and SpliceBERT-human, while similar, is not identical. Notably, RNABERT, despite being 200 times smaller than RNAFM and RNA-SM, achieved a similar performance. This may be due to the pre-training data: RNABERT's dataset is more focused on human intronic sequences, while RNAFM and RNA-SM are more diverse, which could be detrimental to their performance.

The fixed 3' splice site in our dataset simplifies BP detection, as reflected in the baseline model's performance. However, these models rely on prior knowledge of the exon-intron boundaries in a given sequence. Generalizing BP detection to sequences without known splice sites can benefit biological applications. In such cases, models must be trained on arbitrary sequences containing both exons and introns, with the BPs at varying positions. While it is unclear how our models would perform on such data, their substantial improvement over the baseline suggests that they have learned complex patterns beyond the average BP positions. Future work should explore why certain models outperform others by performing a deeper analysis of learned representations. This may provide insights into the specific features or sequence motifs leveraged by high-performing models like ERNIE-RNA and SpliceBERT. Additionally, further investigation is required to analyze whether the improved performance could be due to data leakage from pre-training, or sequence homology between the pre-training and training datasets, as these factors can contribute to overfitting.

5 Conclusion

Some of our fine-tuned RNA language models demonstrate a notable improvement in branch point prediction over previous state-of-the-art models. In particular, Ernie-RNA and SpliceBERT achieved

AP scores of 0.868 and 0.869 respectively, outperforming the fine-tuned model from the original SpliceBERT paper. This improvement is likely due to the size and diversity of our dataset, which allowed for more effective fine-tuning. Moreover, our results suggested that longer training times and further hyperparameter tuning could yield additional improvements.

These models have practical applications for early disease detection through in-silico mutagenesis by identifying how changes in intronic sequences alter BP position. In addition, our study shows the potential of using smaller models in resource-constrained settings without substantially sacrificing performance.

Overall, this work establishes a new benchmark for branch point prediction and suggests several directions for further research. Potential avenues include more extensive hyperparameter tuning and exploring alternative data augmentation strategies. Additionally, a deeper study of learned representations could provide insights into why some models outperform others, and examining potential data leakage or sequence homology from the pre-training datasets may help clarify whether these performance differences are due to overfitting. These efforts will deepen our understanding of RNA splicing mechanisms and support the development of more robust diagnostic tools for genetic disorders.

DATA AND CODE AVAILABILITY

The data used in this project, along with the weights and biases of the fine-tuned models, are accessible on Hugging Face. The code for fine-tuning, testing of models, and generating the plots, is available on GitHub. Model training and experimentation were conducted using NVIDIA A100 GPUs through Google Colab.

ACKNOWLEDGMENTS

This work was carried out as an extension of a course project for EPFL's CS433 Machine Learning class, in collaboration with the Fellay Lab - Human Genomics of Infection and Immunity. One of the authors, Pablo Rodenas Ruiz, is supported by a fellowship from the *la Caixa Foundation* (ID 100010434), with fellowship code LCF/BQ/EU23/12010085.

REFERENCES

Manato Akiyama and Yasubumi Sakakibara. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, 4 (1):lqac012, 02 2022. ISSN 2631-9268. doi: 10.1093/nargab/lqac012. URL https://doi.org/10.1093/nargab/lqac012.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. Self-supervised learning on millions of primary rna sequences from 72 vertebrates improves sequence-based rna splicing prediction. *Briefings in Bioinformatics*, 25(3):bbae163, 04 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae163. URL https://doi.org/10.1093/bib/bbae163.

Zhiyuan Chen and Sophia Y. Zhu. Multimolecule, May 2024. URL https://doi.org/10.5281/zenodo.12638419.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. A 5' utr language model for decoding untranslated regions of mrna and function predictions. *bioRxiv*, 2023. doi: 10.1101/2023.10.11.561938. URL https://www.biorxiv.org/content/early/2023/10/14/2023.10.11.561938.

Suzanne Clancy. Rna splicing: Introns, exons and spliceosome. *Nature Education*, 1(1):31, 2008. © 2008 Nature Education.

Laura De Conti, Marco Baralle, and Emanuele Buratti. Exon and intron definition in pre-mrna splicing. *WIREs RNA*, 4(1):49–60, 2013. doi: https://doi.org/10.1002/wrna.1140. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1140.

Andrew G. L. Douglas and Matthew J. A. Wood. Rna splicing: disease and therapy. *Briefings in Functional Genomics*, 10(3):151–164, 05 2011. ISSN 2041-2649. doi: 10.1093/bfgp/elr020. URL https://doi.org/10.1093/bfgp/elr020.

Ali Saadat and Jacques Fellay. Dna language model and interpretable graph neural network identify genes and pathways involved in rare diseases. In *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pp. 103–115. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.langmol-1.13. URL http://dx.doi.org/10.18653/v1/2024.langmol-1.13.

Ali Saadat and Jacques Fellay. From mutation to degradation: Predicting nonsense-mediated decay with nmdep, 2025. URL https://arxiv.org/abs/2502.14547.

Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, pp. 1–12, 2024.

Jiuyong Xie, Lili Wang, and Ren-Jang Lin. Variations of intronic branchpoint motif: identification and functional implications in splicing and disease. *Communications Biology*, 6(1):1142, 2023.

Weijie Yin, Zhaoyu Zhang, Liang He, Rui Jiang, Shuo Zhang, Gan Liu, Xuegong Zhang, Tao Qin, and Zhen Xie. Ernie-rna: An rna language model with structure-enhanced representations. *bioRxiv*, 2024. doi: 10.1101/2024.03.17.585376. URL https://www.biorxiv.org/content/early/2024/03/17/2024.03.17.585376.

Peng Zhang, Quentin Philippot, Weicheng Ren, Wei-Te Lei, Juan Li, Peter D. Stenson, Pere Soler Palacín, Roger Colobran, Bertrand Boisson, Shen-Ying Zhang, Anne Puel, Qiang Pan-Hammarström, Qian Zhang, David N. Cooper, Laurent Abel, and Jean-Laurent Casanova. Genome-wide detection of human variants that disrupt intronic branchpoints. *Proceedings of the National Academy of Sciences*, 119(44):e2211194119, 2022. doi: 10.1073/pnas.2211194119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2211194119.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, Yonghong Tian, Jian Zhan, Jie Chen, and Yaoqi Zhou. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3–e3, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1031. URL https://doi.org/10.1093/nar/gkad1031.

A Appendix: Baseline Model

The sequences in the dataset have a median of 1558 NTs, see Fig. 3. While this initially suggests that identifying BPs is extremely difficult, the task is more manageable if we take into account the BP distributions. BPs are predominantly adenines (95% in the training set, with cytosine, uracil, and guanine at 3%, 1%, and 1%, respectively) and, as shown in Fig. 4, BPs are typically located within 15 to 40 NTs from the 3' splice site.

These insights allow us to implement a baseline model to evaluate our fine-tuned models. For each NT in a given validation or test sequence, the baseline "model" will multiply the probabilities in Fig. 4 by the per-NT probabilities (e.g. 95% for adenine), resulting in a set of ranking coefficients that we can use to rank the chance of any NT being the BP.

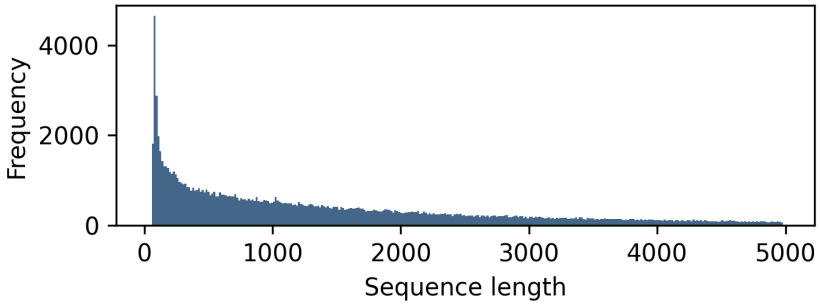

Figure 3: Distribution of sequence lengths in the training set showing 80% of the dataset with the longest sequences excluded.

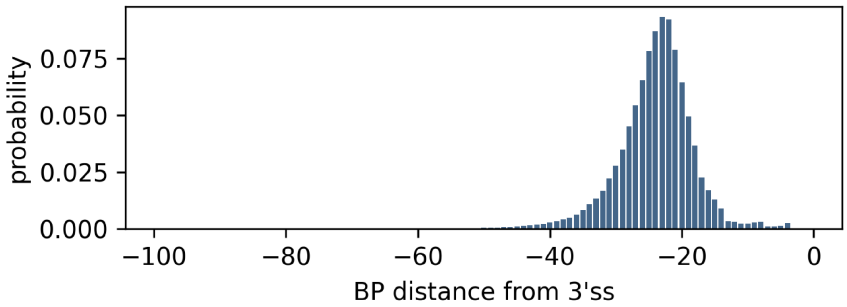

Figure 4: Probability distribution of the distance from the BP to the 3' splice site.

B Appendix: Training and evaluation scheme

The models take as input a NT sequence and output the probability of each NT being the BP. Then, these probabilities are converted into the predicted binary labels according to a decision boundary. All metrics are computed at the NT level, except for sequence accuracy, which is defined as the proportion of sequences for which every NT in that sequence is correctly labeled.

All models were trained using two hyperparameter configurations, presented in Table 3. These hyperparameters were determined through a preliminary, non-exhaustive search, with the training objective of minimizing cross-entropy loss. Based on performance, Set 2 was selected for SpliceBERT and ERNIE-RNA, while Set 1 was used for the remaining models.

Due to the large size of both the dataset and the models, training was computationally expensive. The models were trained on an A100 GPU, taking approximately 3 hours for SpliceBERT and around 12 hours for larger models like RNA-FM (over 3 epochs of the full dataset). Due to time and resource constraints, extensive hyperparameter optimization and cross-validation were not conducted, which is suboptimal as these hyperparameters can significantly influence the models' performance.

Table 3: Hyperparameters used for training the models.

HYPERPARAMETER	SET 1	SET 2
Optimizer	AdamW	AdamW
Learning rate	3×10^{-4}	2×10^{-5}
Weight decay	0.001	0.01
Epochs	3	3
Batch size	16*	16

The decision boundary was optimized using the validation set during training to maximize the F_1 score. Unlike other hyperparameters, optimizing the decision boundary does not require retraining the model, making it computationally inexpensive. To monitor performance during training, metrics on the validation set were computed every 1000 steps. After training, the final evaluation of each model was conducted on the test set.

C Appendix: Performance during Training

The sequence accuracy per training step is shown in Fig. 5. Accuracy grows rapidly during the initial part of the training, reaching a slower but steady growth after the initial spike. These trends demonstrate that these models would benefit from longer training times, with stable gains in performance over time.

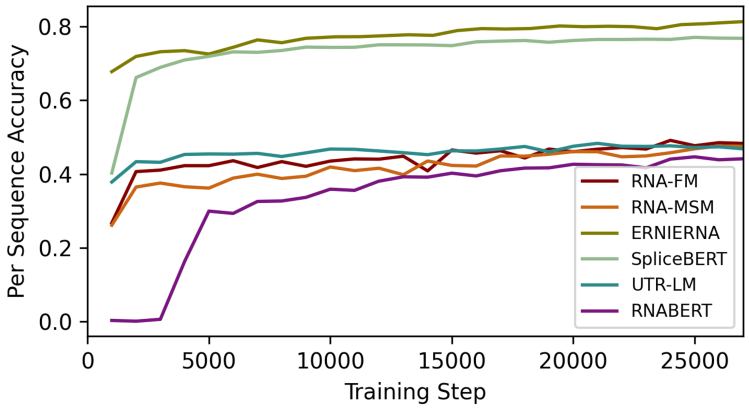

Figure 5: Sequence accuracy during training evaluated on the validation set.

*A batch size of 12 was used for RNA-FM due to VRAM limitations.