Using large scale transfer learning to highlight the role of chromatin state in intron retention

Ahmed Daoud and Asa Ben Hur

Department of Computer Science, Colorado State University, Fort Collins, CO, USA

Abstract

Complex deep learning models trained on very large datasets have become key enabling tools for current research in natural language processing and computer vision. By providing pre-trained models that can be fine-tuned for specific applications, they enable researchers to create accurate models with minimal effort and computational resources. Large scale genomics deep learning models come in two flavors: the first are large language models of DNA sequences trained in a self-supervised fashion, similar to the corresponding natural language models; the second are supervised learning models that leverage large scale genomics datasets from ENCODE and other sources. We argue that these models are the equivalent of foundation models in natural language processing in their utility, as they encode within them chromatin state in its different aspects, providing useful representations that allow quick deployment of accurate models of gene regulation. We demonstrate this premise by leveraging the recently created Sei model to develop simple, interpretable models of intron retention, and demonstrate their advantage over models based on the DNA langauage model DNABERT-2. Our work also demonstrates the impact of chromatin state on the regulation of intron retention. Using representations learned by Sei, our model is able to discover the involvement of transcription factors and chromatin marks in regulating intron retention, providing better accuracy than a recently published custom model developed for this purpose.

Availability: The source code for this work is available at https://github.com/Addaoud/IntronRetention.

1 Introduction

As in natural language processing and computer vision, deep learning has changed the way we analyze large biological datasets, particularly those relating to the complexities of gene regulation in all its aspects [10]. When using deep learning to address a problem of interest you can directly train a network to solve that problem using available labeled data. The other approach, using transfer learning, is to leverage a network trained on related tasks where large training sets are available. This has become the standard approach in natural language processing and computer vision, where a standard set of models have been established as building blocks for downstream applications. Examples

include extremely large NLP models such as BERT [16] and GPT-3 [5] trained in a self-supervised fashion. These are often referred to as emphfoundation models [4]. The same techniques have been applied to train large transformer models on large scale protein and DNA sequence datasets, yielding models that provide useful representations of biological sequences [9, 23, 12]. However, despite their usefulness, these models have not had the level of impact of the foundation models used in NLP. The foundation models of NLP benefit from the availability of extremely large datasets. Compared to those models, the DNA and protein language models have only limited amounts of data. Furthermore, the language of biological sequences is noisier and more difficult to interpret, and is incomplete: its biological function is dependent on chromatin state and the overall cellular state in which it operates. We argue that the information provided by chromatin state profiles and expression values goes towards filling in this missing information and helps deep learning models create more meaningful representations of a DNA sequence.

In the area of gene regulation, another type of large scale model has been developed: one that is trained in a supervised way to predict different aspects of chromatin state. The first such models were DeepSEA [33] and Basset [14]. In Basset, Kelley et al. created a model to predict chromatin accessibility captured as peaks in DNase I-seq data across close to two hundred experiments [14]. DeepSEA included ChIP-seq data for TF binding of 160 TFs across 690 datasets and 104 histone-mark profiles in addition to 125 chromatin accessibility datasets. In both approaches convolutional networks were used to model regions detected as statistically significant peaks in DNase I-seq or ChIP-seq data. These were the early supervised foundation models of chromatin state. The Basenji model [15] took the ideas from the Basset model a couple of steps further, replacing modeling of statistically significant peaks to predicting read coverage in 128-bp bins across the entire human genome. This led to a big jump in accuracy compared to the peak-based modeling employed in Basset, also the result of a shift to the use of dilated convolution, which allows the model to learn shared patterns across larger distances. The Enformer model builds on Basenji and includes attention-based layers that are able to model long-range relationships even better than dilated convolution [2]. The Sei model is the further development of the DeepSea and DeepSea-beluga models [33, 34] and provides predictions of an extensive set of chromatin-state descriptors that include transcription factor binding, chromatin accessibility, and histone modifications for a total of 21,907 genome-wide chromatin-state profiles [7].

In this work we demonstrate the value of foundation models of chromatin state for accurate prediction of human introns that are subject to intron retention (IR). It is well established that splicing is predominantly co-transcriptional [27], and that chromatin state has a strong influence on splicing outcome [20]. In earlier work we have presented a deep learning model that is capable of accurately predicting retained introns in regions of open chromatin, and demonstrated that transcription factors have a significant impact on the regulation of this phenomenon [28]. Since our dataset was based on data generated by the authors of Basset, we explored models based on their architecture; in this case transfer learning had somewhat limited value, providing accuracy that is similar to that obtained by training a model from scratch. In this work we chose to use the Sei model, which is a next generation chromatin foundation model [7]. It is a good match for the task at hand as it models a large number of characteristics of chromatin state, and also uses a

relatively short sequence length compared to models like the Enformer, making it easier to use. By leveraging the knowledge encoded within the pre-trained Sei model, we were able to quickly train models with higher accuracy than our earlier models; furthermore, by directly using the predicted chromatin state variables predicted by Sei, we were able to use a simple logistic regression approach that yielded a highly interpretable model enabling us to make direct connections between transcription factors and their involvement in intron retention and discover a much larger set of transcription factors involved in this process. The methodology used here is similar to the one employed by the authors of the recent EPCOT model [32]. EPCOT is a chromatin state foundation model that has the added feature of using chromatin accessibility as part of its input which provides better generalization across cell-types. They demonstrated the ability of their chromatin state model to perform well in predicting gene expression and enhancer activity.

2 Results and Discussion

In what follows we compare multiple approaches for transfer learning to capture the relation between intron retention and chromatin state, and demonstrate their merit in comparison to our earlier work [28]. Our analysis uses the dataset we created and analyzed for this task, with some additional filtering as described in Section 4. The resulting dataset contains 16,252 DNA sequences associated with intron retention events (positive examples) and 55,888 intronic sequences that are not known to exhibit intron retention based on existing gene models (negative examples). All the examples belong to regions of open chromatin as captured by DNase I-seq data, i.e. are Dnase I hypersensitive sites (DHSs) (see Section 4 for more details).

Transfer learning is all you need. In order to predict intron retention, we repurposed the Sei framework by taking advantage of all its trained convolutional layers. We then applied an additional convolutional layer, followed by an attention pooling strategy, a dense layer, and the output layer as shown in Figure 1. Refer to Section (4) for more details on the choice of the architecture and model training. For comparison, we also trained a model using the DNABERT-2 language model as a building block [35] by adding the same block used to fine-tune the Sei model.

We compared the classification results of all described approaches against a Basset-like model [14] that has three convolutional layers followed by three fully connected layers trained using the pipeline provided by Ullah et al. [28]. Using one hot encoded DNA sequences as input for both architectures, the Fine-tuned Sei achieved an AUPRC of 0.589 compared to 0.52 for the Basset-like model trained from scratch (see Figure 2 for precision recall curves and AUC curves). The Sei-based model also outperformed the LLM-based model that uses DNABERT-2 which had an AUPRC of 0.564. This underscores the effectiveness of using the Sei framework in enhancing predictive accuracy for this task, and that the chromatin representation it learns is highly relevant for the task of predicting retained introns in regions of open chromatin. Its performance advantage over the model trained from scratch likely stems from it having been trained on the entire human genome, giving it the opportunity to more fully learn the language of chromatin-based gene regulation.

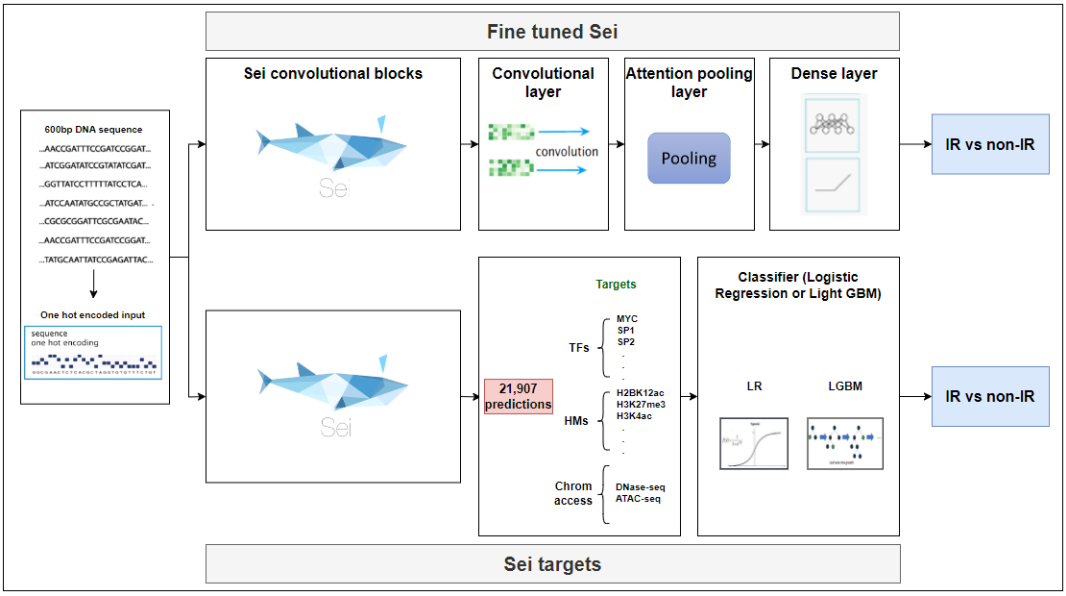

Figure 1: Two approaches for leveraging the Sei architecture for predicting retained introns. Top: The Fine-tuned Sei version uses the representation learned by the Sei convolutional layers, adding an additional convolutional block plus pooling and fully connected and output layers. Bottom: The Sei-outputs version uses the chromatin targets learned by Sei (transcription factor binding, histone modifications, and chromatin accessibility) and uses these as input to a logistic regression or LightGBM classifier.

Interpretable prediction of IR using Sei-targets While the fine-tuned Sei approach described above provides high accuracy, deriving biological insight from it can be challenging. Therefore, we provide an alternative, much simpler approach that is easily interpretable, faster to train, and achieves prediction accuracy that is close to the fine-tuned Sei model. The alternative approach, which we refer to as *Sei-targets*, uses the outputs produced by Sei, which are 21,907 chromatin-state profiles that describe transcription factor binding, open chromatin, and chromatin modifications. These outputs are provided as input features to a simple classifier, either logistic regression or LightGBM trained to predict retained introns based on these 21,907 descriptors of chromatin state provided by Sei (see Figure 1).

The Sei-targets model comes in three flavors depending on which subset of Sei targets were provided as input:

1. All targets: this model uses all 21,907 Sei chromatin profiles as input.

2. TFs model: this model uses the 9,471 predictions associated with transcription factors and other DNA binding proteins were kept when creating this dataset. This dataset includes around 1,000 non-histone DNA-binding proteins.

3. HMs model: this model uses the 10,064 Sei outputs associated with histone marks. This includes 77 different histone targets and marks.

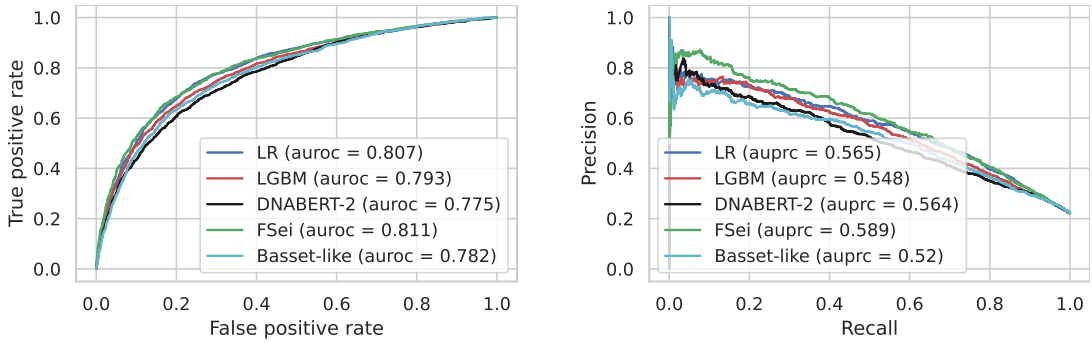

Figure 2: Classification accuracy of deep learning models for prediction of intron retention. These include the trained-from scratch Basset-like architecture, a model that uses embeddings generated by DNABERT-2, an analogous model that uses Sei embeddings (FSei), and two models that use the outputs generated by Sei as features for a simple logistic regression model (LR) and LightGBM (LGBM). We provide the ROC and Precision-recall curves along with the auroc and auprc of each model in the legend.

Based on these three sets of targets we built three logistic regression and three LightGBM models to predict intron retention using these sets of targets as input (see Figure 1). Both classifiers performed similarly well on the TFs dataset, with a slight advantage for the logistic regression classifier (see Figure 2), and in what follows we focus on the logistic regression model. First, we observed that all three logistic regression models had very similar performance, and outperformed the Basset-like model, which had AUPRC of 0.52 compared to 0.564-0.566 for the Sei-targets models (see Figure 3). The more complex fine-tuned Sei model, with an AUPRC of 0.589, had a slight performance advantage. Thus, we conclude that predicted chromatin-state profiles from the Sei framework in conjunction with a with a simple logistic regression approach are capable of predicting retained introns with high accuracy.

We also note that we can obtain the same level of performance using only a subset of the targets: using either the TF targets or the histone modification targets is sufficient. This redundancy between histone marks and TF binding has also been observed in the context of prediction of gene expression [8]. This is not surprising, since histone marks contribute to TF binding [26, 3], and vice versa - TFs help recruit histone modifiers [3]. Finally, an important advantage of the simple logistic regression approach over the fine-tuned Sei is that it is amenable to easy interpretation by looking at the weights assigned to the different targets as discussed next.

TF involvement in IR. To quantify the contribution and involvement of TFs in IR, we considered the weights of the logistic regression model and the feature importance assigned by the LightGBM classifier (see Figure 4). Before discussing the rankings, we note that a given DNA binding protein can occur multiple times on the list, as a given protein might have been assayed in different cell lines or tissues, and therefore associate with multiple Sei targets. And indeed, among the 9,471 targets there are only about 1,100 unique proteins.

We observe many similarities in the rankings produced by the two classifiers, and

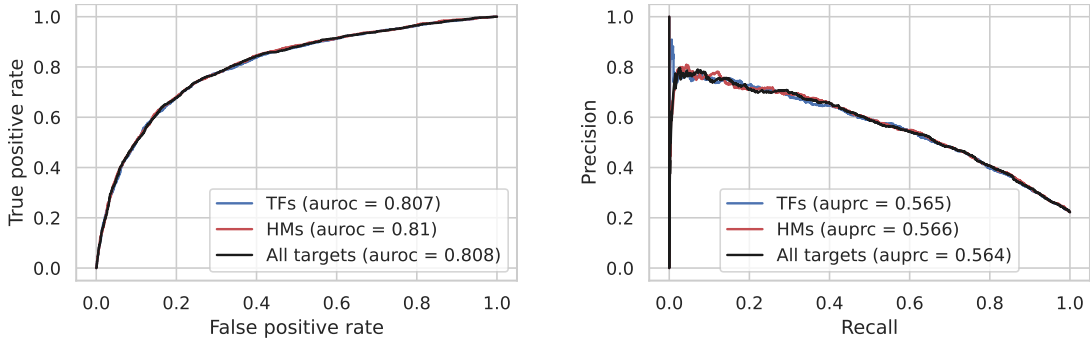

Figure 3: Classification accuracy of the Logistic Regression model using the three different datasets constructed from Sei predictions. We provide the ROC and Precision-recall curves along with the auroc and auprc of each dataset in the legend.

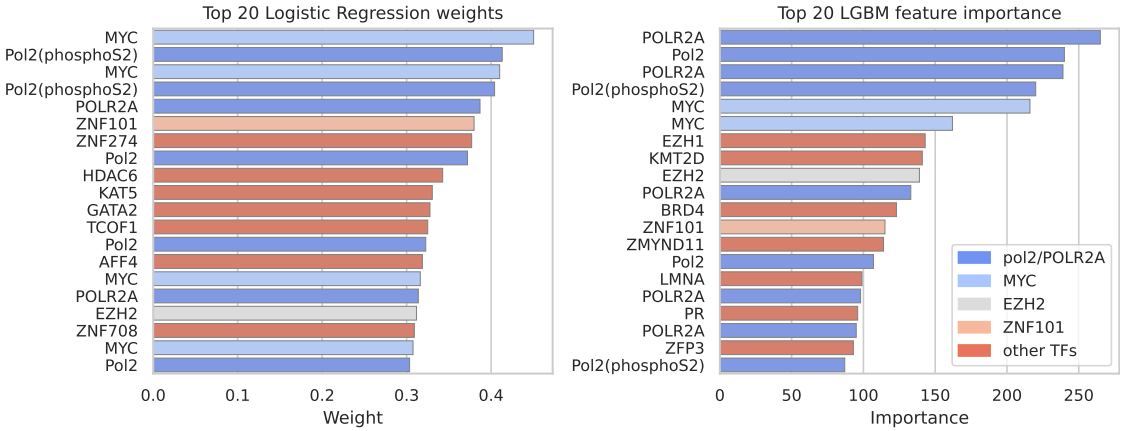

Figure 4: The top 20 targets that contribute to IR in logistic regression (left) and LightGBM (right). Targets are identified by their color.

are able to provide evidence for the involvement of several of the top ranking targets in alternative splicing. First, both classifiers ranked Pol II and modified versions of it (e.g. Pol II modified by phosphorylation of Serine 2) near the top of the list. It is well known that Pol II elongation rate is coupled with alternative splicing, and that modifications of its carboxy terminal domain (CTD) also affect splicing outcome [19]. Therefore it's also not surprising that POLR2A, a major component of Pol II, also appears high on both lists. AFF4 is a core component of the "super elongation complex", and affects the rate of Pol II elongation and hence alternative splicing [22]. MYC is the top ranked TF in both lists, and several studies have linked it to alternative splicing, particularly in relation to cancer [21, 17, 30]. CTCF, is an extremely well studied DNA protein, and is in the top 50 list for logistic regression. It is known to regulate alternative splicing through multiple mechanisms related to its ability to remodel chromatin and affect Pol II elongation [1]. BRD4 was recently shown to regulate alternative splicing [29]. It was ranked 28 by the logistic regression model and 11 by the LGBM classifier

We observed that some TFs have both large positive weights and large negative weights, indicating that a given TF can both promote IR and constitutive splicing in different tissues or cell lines. The top 100 and bottom 100 ranked TFs had an intersection of fifteen TFs that include CTCF, BRD4, POLR2A, and MYC. This is possible since a given TF can be represented in multiple Sei targets describing its activity in different tissues or cell lines.

In our earlier work we have identified six TFs as associated with IR: MAZ, EGR1, SP1, ZBTB7A, SP2, and ZNF263 [28]. Although not at the top of the list, five of them are in the top 5%, and EGR1, is only somewhat lower and had a rank of 606 out of 9,741. Overall, in the top 5% we have 215 different DNA binding proteins with high logistic regression weights. These results suggest that regulation of intron retention is coordinated by a large number of factors, most of which have not been explored until now. The complete list of ranked Sei targets is provided on the github page of this project.

While more difficult, it is still possible to interpret the fine-tuned Sei model and obtain associations of TFs with IR. Using the methodology described in the Methods section, which uses positions identified using the Integrated Gradients method we obtained a ranking of 680 human TFs based on their preference for occurrence in retained introns compared to non-retained introns. The TFs EGR1, SP1, ZBTB7A, and ZNF263 identified by Ullah et al. [28] were all in top 50, while MAZ and SP2 was ranked 86 and 61 respectively. The ranking produced by the Sei-targets approach is not directly comparable to the one produced by Integrated Gradients since each of them is based on a very different set of TFs. Among the two lists (1,097 unique TFs in Sei-targets, and 680 unique TFs in Jaspar), only 334 appeared in both. Thus, the two rankings provide complementary information on the involvement of TFs in IR. We also identified TFs that were strongly associated with non-retained introns, and those likely have a role in constitutive splicing. A complete list is found on the github page of the project. Finally, we note that the Integrated Gradients approach we described can potentially be used to interpret the LLM-based model, although its use of non-overlapping k-mers in creating its embeddings could make this more of a challenge.

3 Conclusions

This study showed the effectiveness of using Sei, a foundation chromatin model, for the downstream application of predicting intron retention, which is known to be co-transcriptionally regulated, and hence impacted by chromatin state. The high accuracy of this approach indicates the large contribution of chromatin state to this phenomenon. The pre-trained model produced superior results compared to building a model from scratch, and also improved on a model based on the DNA language model DNABERT-2. This can be understood from the fact that the Sei model captures more of the complexities of chromatin state due to the much larger datasets used to train it: it was trained on the entire human genome across over twenty thousand experimental genome-wide chromatin-state profiles. Not only that, training was a lot faster than training a model from scratch. Furthermore, the model based on Sei targets was also easy to interpret.

Our results showed that a large number of transcription factors are likely involved in intron retention. However, the mechanism by which this layer of regulation occurs remains to be established. TFs can affect splicing outcome in various ways: either by establishing chromatin state, which in turn affects splicing, or more directly by recruiting splicing factors, and further experiments are required to elucidate the molecular mechanisms of TF regulation of splicing.

4 Methods

Data. Our dataset contains 72,140 DNA sequences of length 600 bp. 16,252 of these are associated with intron retention, and the rest are labeled as non intron retention events. The original dataset was created by Ullah et al. [28]. To prepare the data, they used a collection of DNase1-seq data from multiple cell-lines and tissues with identified DNase I hypersensitive sites (DHSs) [14]. Then, they used SpliceGrapher [24] and iDiffIR [11] to extract IR events from the Ensembl GRCh37 (hg19) reference annotations. The unique IR events overlapping with the DHSs constitute the positive samples, while other intragenic DHSs that did not overlap an IR event were labeled as negative examples. As we were analysing the data, we discovered 495 duplicate sequences, all labeled as Non-IR. Also, using CD-hit, we found 221 sequences that were similar at a threshold of 80%. After removing all these sequences, we were left with 16,231 sequences associated with IR.

To prepare the sequences for use with deep learning, we used one-hot encoding. In one-hot encoding, a sequence is represented as a $4 \times N$ matrix where N is the DNA sequence length. The columns of the matrix represent each position in the sequence assigning a non-zero value at a place corresponding to one of the four DNA nucleotides.

The Sei model. Sei is a deep learning model that predicts the chromatin state of a DNA sequence. It is trained on the entire human genome across many experimental modalities to predict 21,907 genome-wide *cis*-regulatory targets across more than 1,300 cell lines and tissues, including 9,471 TF binding profiles, 10,064 histone mark profiles and 2,372 chromatin accessibility profiles [7]. We split the Sei targets according to the following grouping:

- Targets characterizing chromatin accessibility using DNase I-seq and ATAC-seq across different cell lines and tissues, totalling 2,372 targets.

- Targets associated with histone marks. (10,062 targets).
- The rest of the targets, totalling 9,471 are nonhistone DNA-binding proteins. For simplicity, we refer to them as transcription factors, which constitute the majority.

Sei was trained to predict these profiles from a 4kb sequence represented using one-hot-encoding. The model architecture is composed of three parts: (1) convolutional blocks with dual linear and nonlinear paths; (2) residual dilated convolution layers; and (3) a pooling and classification block. The classification block is one dense layer followed by an output layer of size 21,907.

fine-tuned Sei for predicting IR. This model uses the pre-trained Sei model, repurposing it to predict intron retention. We took the pre-trained Sei convolutional layers up to the B-spline layer. We then applied an additional convolutional layer, followed by an the same attention pooling strategy used in the Enformer model [2]. This is followed by one dense layer and the output layer. We also compared the model performance when freezing the Sei pre-trained weights against fine-tuning them, and we observed that the model yielded better results when we also fine-tuned the pre-trained convolutional layers.

DNABERT-2 for predicting IR. DNABERT-2 is an improved version of the DNA large language model DNABERT, that makes use of an efficient tokenizer and a variety of techniques to get around limits on input length, save time and memory use, with enhanced model performance [12]. DNABERT-2 is a large language model that uses a Transformer Encoder architecture similar to BERT [16]. The authors trained DNABERT-2 with the masked language modeling loss with a mask ratio of 15% on a multi-species genome dataset. The model can be loaded using the Transformers library HuggingFace (Wolf et al. [31]). To leverage DNABERT-2 to predict retained introns we used the embeddings produced by DNABERT-2 as an input to a processing block similar to the one used in the fine-tuned Sei model.

Logistic regression. We used PyTorch to implement the logistic regression model and trained it to predict IR on the basis of Sei targets as input. Additionally, since the data is imbalanced, we used an imbalanced sampler when training the model.

Light Gradient Boosting Machine. LightGBM [13] is a state-of-the-art gradient boosting model with decision trees as the base learning algorithm. It offers the ability to evaluate variable importance, similarly to random forests. Like the logistic regression model, the input to this classifier is the set of Sei targets. In this work we used version 4.0.0 provided by Microsoft, with GPU-acceleration support. We ran the model with the following parameters: $\mathrm{n_estimators} = 1000$ which is the number of trees, $\mathrm{max_depth} = -1$, $\mathrm{num_leaves} = 50$, $\mathrm{bagging_freq} = 1$, $\mathrm{is_unbalance} = True$ which sets the imbalanced mode on, since the data exhibits clear imbalance.

Training, validation, and testing. The data was split into training, validation, and test sets containing 80%, 10%, and 10% of the data, respectively. For all models, we employed a GPU-accelerated training approach. To prevent overfitting and optimize model

performance, we implemented early stopping based on the improvement of the validation loss over a predetermined number of consecutive epochs (the parameter "early stopping epochs" in Table 1). Additionally, to manage computational resources, we set a predefined maximum number of epochs for each model. All the models were trained with the binary crossentropy loss function. The logistic regression, Sei model, and DNABERT-2, were trained using the AdamW optimizer (Adam with weight decay) with the default weight decay parameter. The details of the hyper-parameter space is summarized in Table 1. All experiments were performed on a machine with an NVIDIA GeForce GTX TITAN X GPU with 12GB of memory.

Parameter	Fine-tuned Sei	DNABERT-2	Logistic regression	LGBM	Basset
Maximum epochs	30	20	3000	1000	30
Early stopping epochs	5	5	300	50	NA
Training epochs	30	12	990	950	30
Learning rate	10^{-4}	10^{-4}	10^{-4}	10^{-2}	10^{-3}
Training time (min)	247	240	31	15	5

Table 1: Model hyper-parameters and training statistics. Maximum epochs is the maximum number of epochs allowed for training; after not showing performance improvement for "Early stopping epochs" model training is stopped; the Basset-like model was trained for 30 epochs following Ullah et al. [28]; the actual number of epochs used is provided in "Training epochs". The learning rate was chosen based on an exploratory analysis on the validation set.

Attribution with Integrated Gradients. We propose an approach to score the relevance of each TF using Integrated Gradients (IG) [25], which is an efficient attribution method that assigns an importance score to each position in the sequence. In the first step, we selected all the DNA sequences in the validation and test datasets that were correctly predicted by the fine-tuned Sei with a probability above 0.8. This reduced the number of sequences to 313 IR sequences and 9152 non-IR sequences. Then, we applied Integrated Gradients on the one hot encoded selected sequences to extract all hot spots (positions) that have a gradient value higher than 80% of the maximum value in the same sequence. We used the IntegratedGradients function provided by captum, a model interpretability library for PyTorch [18], to compute the scores. A total of 61,244 hot spots were selected in non-IR sequences and 1,503 in IR sequences. To map the resulting IG scores to TF binding enrichment in IR vs non-IR sequences, we used the Jaspar database [6] to obtain motifs of known TFs. We scored each IG hot spot against the set of Jaspar motifs and kept only those that had a TF score of at least 80% of the maximum score for its motif. For each TF we then scored its enrichment in IR by subtracting the fraction of IR IG hot spots that have a given motif and the fraction of non-IR IG hot spots that have the motif.

References

[1] Adel B Alharbi et al. "CTCF as a regulator of alternative splicing: new tricks for an old player". In: *Nucleic acids research* 49.14 (2021), pp. 7825–7838.

[2] Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature methods* 18.10 (2021), pp. 1196–1203.

[3] Ashley J Bauer and Kathleen A Martin. "Coordinating regulation of gene expression in cardiovascular disease: interactions between chromatin modifiers and transcription factors". In: *Frontiers in cardiovascular medicine* 4 (2017), p. 19.

[4] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[5] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems*. Vol. 33. 2020, pp. 1877–1901.

[6] Jaime A Castro-Mondragon et al. "JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles". In: *Nucleic acids research* 50.D1 (2022), pp. D165–D173.

[7] Kathleen M Chen et al. "A sequence-based global map of regulatory activity for deciphering human genetics". In: *Nature genetics* 54.7 (2022), pp. 940–949.

[8] Chao Cheng and Mark Gerstein. "Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells". In: *Nucleic acids research* 40.2 (2012), pp. 553–568.

[9] Ahmed Elnaggar et al. "ProtTrans: Toward understanding the language of life through self-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* 44.10 (2021), pp. 7112–7127.

[10] Gökcen Eraslan et al. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* 20.7 (2019), pp. 389–403.

[11] Sergei A Filichkin et al. "Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching". In: *Frontiers in plant science* 9 (2018), p. 5.

[12] Yanrong Ji et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome". In: *Bioinformatics* 37.15 (2021), pp. 2112–2120.

[13] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

[14] David R Kelley, Jasper Snoek, and John L Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". In: *Genome research* 26.7 (2016), pp. 990–999.

[15] David R Kelley et al. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks". In: *Genome Research* 28.5 (2018), pp. 739–750.

[16] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of naacL-HLT*. Vol. 1. 2019, p. 2.

[17] Cheryl M Koh et al. "MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis". In: *Nature* 523.7558 (2015), pp. 96–100.

[18] Narine Kokhlikyan et al. "Captum: A unified and generic model interpretability library for pytorch". In: *arXiv preprint arXiv:2009.07896* (2020).

[19] Manuel J Muñoz, Manuel de la Mata, and Alberto R Kornblihtt. "The carboxy terminal domain of RNA polymerase II and alternative splicing". In: *Trends in biochemical sciences* 35.9 (2010), pp. 497–504.

[20] Shiran Naftelberg et al. "Regulation of alternative splicing through coupling with transcription and chromatin structure". In: *Annual Review of Biochemistry* 84 (2015), pp. 165–198.

[21] John W Phillips et al. "Pathway-guided analysis identifies Myc-dependent alternative pre-mRNA splicing in aggressive prostate cancers". In: *Proceedings of the National Academy of Sciences* 117.10 (2020), pp. 5269–5279.

[22] Mahmoud-Reza Rafiee et al. "RNA polymerase II-associated proteins reveal pathways affected in VCP-related amyotrophic lateral sclerosis". In: *Brain* 146.6 (2023), pp. 2547–2556.

[23] Alexander Rives et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e2016239118.

[24] Mark F Rogers et al. "SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data". In: *Genome biology* 13.1 (2012), pp. 1–17.

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[26] Amanuel Tafessu et al. "H3.3 contributes to chromatin accessibility and transcription factor binding at promoter-proximal regulatory elements in embryonic stem cells". In: *Genome biology* 24.1 (2023), p. 25.

[27] Hagen Tilgner et al. "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs". In: *Genome research* 22.9 (2012), pp. 1616–1625.

[28] Fahad Ullah et al. "Evidence for the role of transcription factors in the co-transcriptional regulation of intron retention". In: *Genome Biology* 24.1 (2023), pp. 1–16.

[29] Sheetal Uppal et al. "The bromodomain protein 4 contributes to the regulation of alternative splicing". In: *Cell reports* 29.8 (2019), pp. 2450–2460.

[30] Laura Urbanski et al. "MYC regulates a pan-cancer network of co-expressed oncogenic splicing factors". In: *Cell reports* 41.8 (2022).

[31] Thomas Wolf et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020, pp. 38–45.

[32] Zhenhao Zhang et al. "A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome". In: *Nucleic Acids Research* 51.12 (2023), pp. 5931–5947.

[33] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nature methods* 12.10 (2015), pp. 931–934.

[34] Jian Zhou et al. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". In: *Nature genetics* 50.8 (2018), pp. 1171–1179.

[35] Zhihan Zhou et al. "DNABERT-2: Efficient foundation model and benchmark for multi-species genome". In: *arXiv preprint arXiv:2306.15006* (2023).